COMPUTATIONAL APPROACHES TO STOCHASTIC SYSTEMS IN PHYSICS AND BIOLOGY

BY

PATRICIO RODRIGO JERALDO MALDONADO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Michael Stone, Chair
Professor Nigel Goldenfeld, Director of Research
Professor John Stack
Assistant Professor Yann Chemla

Abstract

In this dissertation, I devise computational approaches to model and understand two very different systems which exhibit stochastic behavior: quantum fluids with topological defects arising during quenches and forcing, and complex microbial communities living and evolving withing the gastrointestinal tracts of vertebrates. As such, this dissertation is organized into two parts.

In Part I, I create a model for quantum fluids, which incorporates a conservative and dissipative part, and I also allow the fluid to be externally forced by a normal fluid. I use then this model to calculate scaling laws arising from the stochastic interactions of the topological defects exhibited by the modeled fluid while undergoing a quench.

In Chapter 2 I give a detailed description of this model of quantum fluids. Unlike more traditional approaches, this model is based on Cell Dynamical Systems (CDS), an approach that captures relevant physical features of the system and allows for long time steps during its evolution. I devise a two step CDS model, implementing both conservative and dissipative dynamics present in quantum fluids. I also couple the model with an external normal fluid field that drives the system. I then validate the results of the model by measuring different scaling laws predicted for quantum fluids. I also propose an extension of the model that also incorporates the excitations of the fluid and couples its dynamics with the dynamics of the condensate.

In Chapter 3 I use the above model to calculate scaling laws predicted for the velocity of topological defects undergoing a critical quench. To accomplish this, I numerically implement an algorithm that extracts from the order parameter field the velocity components of the defects as they move during the quench

process. This algorithm is robust and extensible to any system where defects are located by the zeros of the order parameter. The algorithm is also applied to a sheared stripe-forming system, allowing the calculation of the corresponding scaling laws.

In Part II, I investigate the evolutionary dynamics of communities of microbes living in the gastrointestinal tracts of vertebrates, and ask to what degree their evolution is niche-driven, where organisms fitter to their environment become dominant, or if it is neutral, where the organisms evolve stochastically and are otherwise functionally equivalent within their communities. To that end, a series of computational tools were developed to pre-process, curate and reduce the data sets.

In Chapter 4, I analyze the raw data for this research, namely short reads of 16S ribosomal RNA, and quantify how much of phylogenetic information is lost by using these short reads instead of full-length reads, and show that for lengths spanning 300 to 400 base pairs, we can recover some meaningful phylogenetic information.

In Chapter 5, I introduce a pipeline for pre-processing, alignment and curation of libraries of short reads of rRNA. We show that this pipeline significantly reduces the artifacts usually associated with these sequences, resulting in better clustering of the sequences, and better phylogenetic trees representing their organismal relationships.

In Chapter 6 I use the data processed with the above tools to analyze communities of microbes living in gastrointestinal tracts of vertebrates, and we ask to what extent the evolutionary dynamics of these communities is dominated by niche-based evolution, or if the communities behave neutrally, where evolution is random and all organisms are functionally equivalent. We conclude that there is evidence for strong niche-based dynamics, though we cannot fully discard some degree of neutral evolution.

Finally, in Chapter 7 I propose a method to quantify the balance present in phylogenetic trees to compare a large-scale molecular phylogeny to full organismal taxonomies. I show that there is considerable structure in all of them, but direct comparison of both types of trees is difficult at the moment due to their different intrinsic structure.

To my mother.

Acknowledgements

At the end of this journey, it is hard, or dare I say impossible, to fairly thank everyone who had a hand in making this possible, one way or another. It is also hard for me to fairly express those feelings in written words, but I'll make clear that those feelings do exist, and are very genuine. I hope that in the following I manage to appropriately acknowledge all these people, many of whom would not[1] (and probably should not!) read this dissertation.

For a brand new graduate student, the phrase "interested in everything" is frankly intimidating. After the first few weeks under Nigel's guidance, I certainly realized that this kind of appetite for knowledge is rarer than expected within the sciences, and I also realized how far it can take you. I consider myself very lucky for having been part of Nigel's group all these years. His guidance, kindness, dedication and intelligence have definitely shaped my life in ways I definitely could not imagine the moment I set foot in Loomis lab. I will certainly miss all the frank talks, the long office meetings on the quest to try the latest crazy idea to solve our problems, and the long (maybe too long) arguments about the latest happening in the geek world. I am most thankful, and wish him the best.

An integral part of my experience here were of course my fellow students and postdocs of the group over the years: John Veysey, Badri Athreya, Patrick Chan, Nicholas Guttenberg, Tom Butler, Zhenyu Wang, Maksim Sipos, Hong-Yan Shih, Farshid Jafarpour, Vikiath Rao, David Reynolds, Nicholas Chia, Luiza Angheluta, Andreas Menzel and Michael Assaf. I am thankful for all the ideas shared, all the tips offered to survive the group, all the discussions about everything from bread making to politics to bad science fiction

[1] I couldn't resist the temptation of paraphrasing the dedication in the book by Pismen [1], one of the central pieces of literature relevant to this dissertation, and beautifully written despite such an arid subject matter. Also a footnote must exist in this dissertation.

movies (specially the almost fanatical discussions about computer stuff). I have to specially acknowledge Nick Chia, for all these years of blunt advice offered when I needed it the most, and for trusting me for future endeavors. I look forward to the next years working together at the Mayo Clinic. Also my thanks to Luiza, friend, collaborator and aikido pal. I thank her for opening my eyes pretty much since the moment she landed in Urbana, and for showing me an intensity and dedication that I admit I'm envious about. I wish her the best wherever life takes her.

I am very thankful to Prof. Gustavo Gioia and Prof. Bryan White and their students and postdocs for the wonderful discussions over the years, and for the opportunities I was given. I will not forget them.

I am thankful to the physics crowd, in no particular order, Akbar, Esi, Aruna, Vadas, Mohammed, Ry, Mike Bednarz, Mike Bell, John Koster, Jeremy Tan, and many others who certainly deserve to be here. I'm thankful for the journey shared, for the too many long nights spent in the Lab, for the nights spent trying to forget the Lab, and for making this place more bearable. You won't be forgotten, and wish you a successful life.

I am most grateful to the people at Central Illinois Aikikai, and its dojo cho, Knut Bauer Sensei. I am thankful for reviving my passion for Aikido, for creating and keeping such a wonderful practice place in such a little town. I will make sure to pass forward everything I learned there. Also, a very special thanks to Vasi Crecea. Simply not enough words to fit in here, even after all that has been said. Just... thank you. You know why. My thanks also to Dean, friend and roommate, for his kind offering of sanity (and cookies!) when I needed it.

My thanks to the Chilean community in central Illinois, for bringing a taste of home away from home. I'll be sure missing people here, but I must mention David, Judith, Juan Ignacio, Rodrigo, Roberto, the Lepeleys. I thank you for all these years, and for helping me in really dark times. My thanks to Marcos Sotomayor, for his invaluable help and advice to get me jumpstarted in UIUC. My big thanks to Matias Negrete, friend, roommate and confident, for all these years of friendship.

My thanks to the Chicago crowd, Carlos, Cota, for their hospitality and eagerness to have fun. The oh so big Castle crowd, Cesar, Mehmet, Stan, and so many others. Thank you for the insanity when I needed it. The Beauchef crowd, spread thin all over the world, Jorge, Jaime, Karen, Jessica, Lucho, Pancho. Thanks for always cheering, even from a distance. My very special thanks to Felipe, for all these years of friendship, for the hospitality in Chicago, and for the enormous kindness and generosity. I'm sure we'll meet again.

Last, but not least. My family. My mother, Maria Teresa, for her love, kindness and support, even now when I hope she is in a better place than this Earth. My father, Raul, for his always unconditional support and love. My siblings, Lorena and Raul, for their support and for always offering a home away from home. This paragraph will never do justice to all that is needed to be said here. I am very thankful.

It is my hope this thesis honors the memory of my mother.

The work presented in this dissertation was partially supported by the L.S. Edelheit Family Biological Physics Fellowship.

Table of Contents

Chapter 1

Introduction

This thesis is in two parts, one on superfluids and topological defects, the second concerning interacting microbial communities. The work was performed in Nigel Goldenfeld's group, which emphasizes training students in both the methods of theoretical condensed matter physics as well as introducing students to appropriate problems in biology. My research has followed this pattern, developing my skills in quantitative and novel approaches to spatially-extended dynamical systems, followed by innovative approaches to computational problems in evolutionary ecology. The sets of problems that I focused on are difficult in their respective fields because they are dominated by collective effects. The goal of my research has been to devise novel computational approaches to these problems, and as the reader will see, they are necessarily very different; however, the unifying feature is the *ab initio* approach to modeling and analysis that I have developed with my collaborators and mentors. In this chapter, I give an overview of the thesis, and describe my specific contributions to the various projects in which I have participated.

1.1 Overview of projects

During the last two decades, there has been an ongoing effort by the condensed matter and fluid dynamics communities to fully characterize complex and turbulent flow in quantum fluids. This effort has a strong experimental component, where techniques originally developed for traditional fluid mechanics were being applied to superfluids, and after the experimental realization of Bose-Einstein condensates, a whole new avenue of study was opened with an unprecedented degree of control and at extremely low temperatures, albeit in small sizes. Despite these advances, most of the interesting ranges for experiment are still away

from realization. On the theoretical side of the coin, advances have also been made on understanding the dynamics of topological defects in a tangle (which is a description of a turbulent superfluid). In a happy coincidence, the the ascent of unprecedented computer power as a cheap commodity has enabled researchers to simulate these fluids with previously unattainable degree of detail and complexity, allowing questions to be asked in regimes still away from experimental realization. Both experimentalists and theoreticians are asking the following questions about quantum-turbulent systems: what are the similarities and differences between quantum turbulence and normal fluid turbulence? To what degree can we understand this quantum turbulence state by only looking at its topological defects (vortices)? What is the role of excitations in this state? Are there other defect-dominated systems (like crystal plasticity) that can inform about the dynamics of quantum fluids (or vice versa)?

My contribution, detailed in Part I of this dissertation, falls firmly in the camp of numerical research on quantum fluids. I propose an algorithm to efficiently simulate quantum fluids. Although inspired after differential equation-based models, this algorithm is cellular (map-based), allowing us to fully include the relevant physics and permitting a rapid time evolution of the system, which allows for measuring of scaling laws present in the system. I validate this algorithm by calculating known scaling laws and showing dynamics that quantum fluid motion should exhibit, such as quantized vortices, complex vortex structures under external driving. Finally I propose an extension of this model to fully include and couple excitations to the system.

Moving over to a different topic, there is nothing short of a revolution happening in microbiology. Long gone are the days where microbes were only seen as a nuisance, no more than invisible disease vectors. Also on the way out is the disturbing and egotistical thought that microbes are just relics of millennia past, that gave way to superior forms of life such as ourselves, or are, at most, unremarkable and not worthy of study. It was the work of physicists that allowed us, for the first time, to really have a look inside cells and ask questions such as how do genes really work? Shortly thereafter, this approach had a resounding success with the discovery of the Archaea by a physicist [2]: a different form of cellular life was hiding within these "unremarkable" microbes. So different, that it can be considered closer to Eukaryotes (such as ourselves) than to Bacteria, and had the consequence that the Tree of Life had to be redrawn. As technology progressed, along came the dawn of the genomics era. This put in the spotlight the mind-boggling diversity of metabolism and behavior present in microbes. The happy coincidence of the advances in computing

permitted the study of not only single microbes but communities of them. We can now study the complex interactions between them and with their environments, the extent that microbes affect the rest of life on Earth, including ourselves. We can ask questions about their ecology. In this regard, the advantage lies in the large samples we can actually study (millions of microbes as opposed to hundreds or thousands of individuals, for example plants), and this plethora of data is a blessing for us researchers interesting in modeling this behavior. It is also a curse, because it becomes a problem to handle this data, for both curation and analysis.

My contributions described in Part II reflect the path taken from massive amounts of raw data to questions about the nature of evolution within a community of bacteria. My involvement in microbial ecology came just as the field began the transition to massively high-throughput sequencing of communities using metagenomics techniques, such as pyrosequencing. Thus much of my work was devoted to technique development, but we also were able to address a fundamental question about evolutionary ecology. First, I characterize the extent to which we can rely on short fragments of a gene for phylogeny analysis of communities. Then we propose a pipeline to process this raw sequence data from an error-ridden set to clean, curated and aligned sequence libraries. Finally, using this cleaned data we ask questions about evolution models in communities of bacteria living inside the gastrointestinal tracts of vertebrates, to what extent their evolutionary dynamics show specialization in niches, or if all organisms are otherwise functionally equivalent in the community, and how much we can answer using the genomic data at hand.

1.2 Overview of superfluid hydrodynamics

One of the hurdles in the numerical study of quantum fluids is the sheer complexity of simulating a complete, fully coupled fluid including both the condensate and excitations. Recent advances have allowed detailed studies of zero-temperature fluids [3], which have moved from rather artificial and inefficient direct simulation of the quantum vortices (either directly integrating the Biot-Savart model for vortex strings or using a Local Induction Approximation), to fully pseudospectral integrations of the Gross-Pitaevskii equations [4]. All of these approaches lack a fully coupled model for the back-reaction from excitations, though attempts have been made.

1.2.1 Efficient algorithm for simulation of superfluids

In Chapter 2 I propose a model for the superfluid dynamics that can be efficiently integrated using off the shelf computing equipment. This model uses a Cell Dynamical Systems (CDS) approach to account for the physics of the condensate. As such, it is not an integration of differential equations representing the dynamics of the condensate, but directly translates the symmetries and dynamics of the superfluid into computationally-efficient coupled maps that replace those which would arise from a conventional discretization of the governing partial differential equations. Such maps can yield results that give experimental predictions of universal quantities, such as scaling functions, with no adjustable parameters [5–7]. Moreover, in some circumstances it has been possible to compare CDS methods with conventional numerics, and the results clearly indicate that both are in the same universality class [8]. Although the model does not come directly from a differential equation, it is inspired by the two-fluid model from Ginzburg and Pitaevskii [9, 10]. In this chapter I show validation as well as limitations of the model, in the form of scaling laws and behavior of the vorices in the presence of external forcing. I also introduce a numerical implementation of a topological defect tracking algorithm, based on a theoretical representation originally due to Halperin and Mazenko [11, 12]. This algorithm extracts the velocity of topological defects described by zeros of the order parameter, and also extracts their topological charge. This method allows us to calculate the probability distribution of these velocities, helping further validation of the superfluid model. Finally, I propose a model for the excitations, where they are described as a normal fluid modeled using a Lattice Boltzmann [13, 14] approach. The condensate and the excitations are then coupled via advection of the condensate due to the normal velocity field and the condensate affects the normal fluid through its pressure tensor.

1.2.2 Defect velocity distributions after a quench

As an example of the use of the numerical scheme, in Chapter 3 I perform a simulation of a critical quench of an $O(2)$ model in two and three dimensions. Using the defect tracking algorithm I extract the velocity of the defects, as the quenches evolve in time, and compute their probability distribution function. The results show that the high-speed tails of the distributions of velocity v follow a predicted power-law scaling of the form v^{-3}, with a Gaussian correction due to finite-size effects (the defect core has a non-zero size). This scaling reflects the existence of long-range interactions between the individual vortices. A separate example is studied for defects in a stripe-forming system, where defect motion is anisotropic, finding that one degree of

freedom exhibits the power-law scaling, whereas the other degree of freedom shows a Gaussian distribution of velocities. This work was performed in collaboration with Luiza Angheluta.

1.3 Overview on microbial ecology

In Part II, I tackle a problem on evolutionary dynamics where we ask if purely stochastic dynamics can explain the species abundance patterns seen in microbial communities in vertebrates, or if niche selection is still a preferred evolution mechanism. Along these chapters, I start with the preprocessing of raw data. In this case, the raw data are sequences from the gene of the small subunit of ribosomal RNA, which for Bacteria and Archaea is commonly referred as 16S rRNA. This gene has been a favorite tool as a tag for microbial organisms due to its ease of sequencing, its seemingly consistent length, rate of evolution and apparent lack of horizontal gene transfer that complicates phylogenies. I show that despite these ideal qualities, real experimental data still has some problems that need to be attended to before using it to ask relevant biological questions. I then proceed to introduce a pipeline to clean the sequences of these artifacts, and then I proceed to use the data for analysis of evolutionary dynamics. Finally, I present the technical aspects of a related problem, which shows methods for the quantification of the structure of trees as a possible way to compare methods of systematization of phylogenetic and taxonomic data.

1.3.1 16S rRNA sequence data

During the research detailed in this dissertation, the starting point (at least for us in these collaborations) was to process many libraries of 16S rRNA reads into usable information, either as a set of scores characterizing its diversity and composition, or as phylogenetic trees describing its evolutionary relationships. However, limitations arise due to the inaccuracies inherent in the contemporary sequencing platforms, which lead to misreads and spurious data. Another strong limitation is that we were not dealing with a whole read of the 16S rRNA gene, but rather a subset of it (the short reads). Naturally, this leads to a bound in the amount of information that we can in principle extract using these short reads. Previous studies have characterized the effects of these short reads on the different diversity metrics, and also on the determination of taxonomic composition as compared to full-length reads of the gene.

In Chapter 4, I extend the characterization to quantify the information loss on phylogenetic information when short reads of 16S rRNA are used instead of full length reads. This kind of limitation is of high interest,

given that the very rapid evolution of sequencing platforms and its use in environmental microbiology have resulted in platforms with either very high throughput, high quality but even shorter read length (e.g. Illumina HiSeq sequencers), or very long read length but high error rate (e.g. PacBio RS sequencers).

In Chapter 5, I introduce a pipeline, dubbed TORNADO, designed to process 16S rRNA libraries starting from raw sequences into a state suitable for analysis. Developed together with Maksim Sipos and Nicholas Chia, this pipeline takes raw 16S rRNA sequences, removes known artifacts resulting from the pyrosequencing process, leaving them ready to be aligned using both the NAST algorithm [15] and RDP's [16] Infernal aligner [17]. Then we combine both multiple sequence alignments, with the premise that we obtain the best of both worlds by taking the good secondary structure-based alignments of conserved regions aligned using Infernal, and the difficult to align hypervariable regions as aligned using NAST. TORNADO also provides a tool set for semi-automated curation of the alignment to increase its quality. Finally, TORNADO provides a complete linkage-based algorithm to cluster the sequence libraries into Operational Taxonomic Units.

1.3.2 Neutral versus niche evolution as seen with ribosomal RNA sequence data

As a direct application of the above chapters, in Chapter 6 we ask a more basic question in microbial ecology. From an ecological snapshot of a community, as reflected in the abundance distributions of its constituents, what can we infer about the dynamical processes that shaped the community? The two endmembers of such processes are deterministic evolution of populations in response to environmental gradients and stochastic evolution of populations where all species are regarded as functionally equivalent, and thus experience no specific attraction to environmental niches. Our project was to generate high resolution surveys of complex microbial communities, using metagenomics techniques to obtain vast libraries of environmental short reads of 16S rRNA. We developed novel methods to extract process information from these metagenomic data. Starting from libraries of microbial communities from the gastrointestinal tracts of three vertebrates, we first fit the rank-abundance curves of these communities using the predicted curves from Hubbell's neutral theory. Then we contrast this information with evidence for niches as seen in the taxonomic composition of these communities. Finally, we develop a model of niche-based and neutral evolution as we would expect it to see in sequence space, and test this model with the actual sequence data. We conclude that although the rank-abundance curves fit a neutral model of evolution, sequence data suggests a dominant niche-based assembly process.

1.3.3 Balance on phylogenetic trees

To conclude, in Chapter 7 I propose a method to quantify the balance present in phylogenetic trees to compare a large-scale molecular phylogeny to full organismal taxonomies. I show that there is considerable structure in all of them, but direct comparison of both types of trees is difficult at the moment due to their different intrinsic structure.

1.4 My contributions

For the work in Chapter 2, I expanded on a model originally devised by Maurizio Mondello and Nigel Goldenfeld by adding an explicit external driving velocity field (i.e. a normal fluid) using a simple advection scheme. I also created a complementary model for the conservative part of a two-fluid superfluid model, and performed validation of said models. I also proposed a method to fully couple this superfluid model with another model for the normal fluid, via a Lattice Boltzmann model, with an appropriate pressure tensor constraint which implements the back-reaction and closes the coupling between the condensate and excitations. Together with Luiza Angheluta, I helped translate a topological defect tracking algorithm, created by Halperin and Mazenko, from a purely analytical tool to a numerical scheme capable of extracting the velocities of the defects.

For the work in Chapter 3, I performed the numerical simulations for the $O(2)$ models in two and three dimensions. Together with Luiza Angheluta, I performed the data reduction, analysis and interpretation of the results. All the work on the dislocation model under shear flow was done by Luiza Angheluta.

In the work of Chapter 4, I wrote the script to extract randomized sequence data from the Greengenes [18] database, I created the simulated pyrosequencing libraries, I performed the multiple sequence alignment of all libraries and calculated the phylogenetic trees using said alignments. I also coded the Pearson Correlation (PC) comparison metrics and performed the analysis on the different trees using the PC metric, and the two Robinson-Foulds metrics [19], interpreting the results in collaboration with Nicholas Chia.

For the work of Chapter 5, I wrote the script to merge the multiple sequence alignments and, together with Maksim Sipos, we wrote the pre-processing scripts and the web interface for the TORNADO pipeline. All the benchmarking and clustering work was performed by Maksim Sipos. Together with Ani Qu, I helped create and curate the chicken caecum data used in this chapter.

For the work of Chapter 6, together with Maksim Sipos we pre-processed the 16S rRNA sequence data. Then I extracted the taxonomic information for all the different libraries. Together with Maksim Sipos, Nigel Goldenfeld and Bryan White we interpreted the results of the distance model. The metric that allowed us to distinguish niche and neutral dynamics was first suggested by Maksim Sipos after numerous other metrics had been suggested and tried by Sipos, Chia, Nigel and me.

For the work on Chapter 7, I implemented and then applied the tree metrics on the phylogeny and the taxonomies, and then processed the results to make them suitable for visualization.

1.5 List of publications

The work presented in this dissertation has been published or is under review in the following publications:

- L. Angheluta, P. Jeraldo and N. Goldenfeld, *Anisotropic velocity statistics of topological defects under shear flow*, Physical Review E **85**, 011153 (2012)

- P. Jeraldo, N. Chia and N. Goldenfeld, *On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys*, Environmental Microbiology **13**, 3000-3009 (2011)

- M. Sipos, P. Jeraldo, N. Chia, A. Qu, A.S. Dhillon, M.E. Konkel, K.E. Nelson, B. White, and N. Goldenfeld, *Robust Computational Analysis of rRNA Hypervariable Tag Datasets*, PLoS ONE **5**, e15220 (2010)

- P. Jeraldo, M. Sipos, N. Chia, J.M. Brulc, A.S. Dhillon, M.E. Konkel, C.L. Larson, K.E. Nelson, A. Qu, L.B. Schook, F. Yang, B.A. White and N. Goldenfeld, *Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes*, under review at PNAS (accepted by referees)

Part I

Dynamics of Topological Defects

Chapter 2

Complex quantum vortex dynamics in superfluids

We introduce a minimal model for the dynamics of a simple superfluid. This model contains two components: (1) a cell dynamical systems model of the superfluid parameter, and (2) a Lattice Boltzmann scheme for the conservation and evolution equations of the normal fluid component, which represents the excitations of the order parameter field. The model features a natural inclusion of topological defects, annihilation and reconnection of said defects, and strong coupling of the superfluid fraction with its normal counterpart. Preliminary results include verification of the scaling law in vortex reconnection and reproduction of instabilities in vortex line length due to normal flow vorticity. Finally, we propose studies on superfluid turbulence which include detection of vortex reconnection avalanches and characterization of the interaction between dissipation and the turbulent vortex tangle.

2.1 Introduction

Complex quantum flows have recently gained some attention, as longstanding problems are being tackled thanks to developments in theoretical, experimental and numerical areas. In quantum fluids, a restriction on the nature of their flows arises because of its coherent state: its flow must be irrotational on a simply connected domain. Quantum vortices allow for the existence of rotation in superflows as they are a topological defect. Moreover, the vortex core size is constant and fluid circulation about it is *quantized*. A complex

superflow can be characterized by the state of its vortices.

In quantum turbulence, the quantum vortices form a tangle [20]. This tangle is a dynamic state, as the vortices are stretched, undergo reconnections, and rings are created and destroyed. These quantum vortices also interact with their environment. They interact with the gas of excitations (the *normal* fluid) by emitting phonons when dissipating energy, and also get advected by this normal flow. Yet this interaction, and its influence in quantum turbulence, is not completely understood.

For example, the striking similarities between the turbulent regimes of both classical and quantum fluids have led to questions about the nature of energy dissipation in superfluids at very low temperature. Also remarkable are turbulent phenomena unique to superfluids, namely counterflow turbulence, due to helium II's peculiar heat transport properties. This curiosity has trickled down to other related problems, such as complex vortex structures in Bose-Einstein condensates (BEC), and even to more distant areas such as neutron star hydrodynamics. It is interesting to see how far the analogy between classical and quantum turbulence goes, as it is to know where to place the boundaries.

If we consider the problem of complex superflows from a more abstract perspective, we can frame it in terms of the dynamics of topological defects, how they interact with external fields and amongst themselves. We can now relate the problem of complex superflows to seemingly unrelated phenomena such as dislocation avalanches in plastic deformation. And to go even further, the possibility of extending these studies to the rich gamut of states present in superfluid ^{3}He just adds momentum to the community of researchers.

The purpose of Part of this dissertation is to study complex superflows, to obtain insight on the inner workings of the energy dissipation mechanisms present in turbulent superflows, and how they affect the dynamics of the quantum vortices, to characterize the transport of complex scalar fields in a statistically turbulent flow and to probe for the existence of vortex reconnection avalanches in turbulent superflows, analogous to their plastic flow counterparts. Also, in this Part, I develop and adapt numerical techniques to permit the aforementioned studies, and understand their range of validity.

2.2 Background on superfluids

The goal of this section is to introduce a model of superfluid hydrodynamics that is a generalization of the much better known Non Linear Schrödinger Equation (NLSE) model, but valid at $T > 0$. This model will be related to the physics introduced in my own work in section 4. Before describing superfluid, or quantum,

turbulence, it is necessary to do a brief survey on superfluids and the models used to describe them. An ideal superfluid is characterized primarily by the lack of viscosity, entropy and irrotational flow on simply connected domains, all characteristics being quantum in origin. The only way rotation can be achieved in a superfluid is due to the presence of vortices, and circulation about these vortices is quantized. with the quantum of circulation has the value $\kappa = h/m$, with m being the mass of the atom of the element in the superfluid state.

2.2.1 Analytical models for superfluids

2.2.1.1 The nonlinear Schrödinger equation

In superfluid helium the non-superfluid fraction becomes negligible at temperatures below 1 K, and a relatively simple qualitative model can be made. Such a model is the nonlinear Schrödinger equation (NLSE),

$$i\partial_t\psi = -\frac{1}{2}\nabla^2\psi - \left(1 - |\psi|^2\right)\psi \tag{2.1}$$

where ψ is a complex scalar order parameter. This order parameter can be related to physical quantities through the Madelung transformation $\psi = \rho_s^{1/2} e^{i\phi}$. Applying this transformation to equation 2.1 and defining the superfluid velocity $\boldsymbol{v}_s = \nabla\phi$ we find

$$\partial_t\rho_s + \partial_j(\rho_s v_{sj}) = 0 \tag{2.2a}$$

$$\rho_s\partial_t v_{sj} + \rho_s v_{sk}\partial_k v_{sj} = -\partial_j p + \partial_k S_{jk}, \tag{2.2b}$$

where $p = \frac{1}{2}\rho_s$ is the pressure and $S_{jk} = \frac{\rho_s}{4}\partial^2_{jk}\ln\rho_s$ can be thought of a quantum stress. This set of equations resembles the dynamics of a Euler fluid, except for the last term in equation 2.2b. The NSLE equation, despite being valid only in the $T \to 0$ limit, has been very successful at providing insight on different phenomena in simple superfluids, including turbulence.

But, as the temperature is increased, the effect of the thermal excitations cannot be neglected anymore, and a different framework must be used.

2.2.1.2 Landau and Tisza's two fluid model

At intermediate temperatures in between $T = 0$ and $T = T_c$, the superfluid is composed of the ideal superfluid fraction and the gas of thermal excitations (phonons and rotons). Between late 1930s and early 1940s, Landau and Tisza proposed a *two fluid model* of superfluid helium [21]. In this model, the superfluid is modeled as two interpenetrating fluids, one being the superfluid, which is inviscid, irrotational and carries no entropy, the other one being an otherwise normal, classical fluid. Each of the fluids is described by its own density and velocity fields. The equations for the two fluid model are

$$\rho_{\mathrm{s}} \frac{\mathrm{d}\boldsymbol{v}_{\mathrm{s}}}{\mathrm{d}t} = -\frac{\rho_{\mathrm{s}}}{\rho}\nabla p + \rho_{\mathrm{s}} S \nabla T + \frac{\rho_{\mathrm{n}}\rho_{\mathrm{s}}}{2\rho}\nabla(\boldsymbol{v}_n - \boldsymbol{v}_s)^2 - \boldsymbol{F}_{\mathrm{ns}} \tag{2.3a}$$

$$\rho_{\mathrm{n}} \frac{\mathrm{d}\boldsymbol{v}_{\mathrm{s}}}{\mathrm{d}t} = -\frac{\rho_{\mathrm{n}}}{\rho}\nabla p + \rho_{\mathrm{n}} S \nabla T - \frac{\rho_{\mathrm{n}}\rho_{\mathrm{s}}}{2\rho}\nabla(\boldsymbol{v}_n - \boldsymbol{v}_s)^2 + \boldsymbol{F}_{\mathrm{ns}} + \eta\nabla^2\boldsymbol{v}_{\mathrm{n}}, \tag{2.3b}$$

with the constraint $\nabla \times \boldsymbol{v}_{\mathrm{s}} = 0$ and the conservation laws

$$\partial_t \rho + \nabla \cdot \boldsymbol{j} = 0 \tag{2.3c}$$

$$\partial_t(\rho S) + \nabla \cdot (\rho S \boldsymbol{v}_{\mathrm{n}}) = 0. \tag{2.3d}$$

Here, ρ_{s} and ρ_{n} are the densities of the superfluid and normal fractions, respectively, $\boldsymbol{v}_{\mathrm{s}}$ and $\boldsymbol{v}_{\mathrm{n}}$ are the corresponding velocity fields, $\rho = \rho_{\mathrm{s}} + \rho_{\mathrm{n}}$ is the total density of the fluid, $\boldsymbol{j} = \rho_{\mathrm{s}}\boldsymbol{v}_{\mathrm{s}} + \rho_{\mathrm{n}}\boldsymbol{v}_{\mathrm{n}}$ is the total mass current, T is the temperature and S is the entropy. It must be noted that equation 2.3d is valid only when dissipation is small.

The two fluids are described by Navier-Stokes-like equations, but they are coupled via a *mutual friction* force (the term $\boldsymbol{F}_{\mathrm{n_s}}$), a phenomenological parameter describing the interaction between the two components. Its form, unfortunately, depends on the problem being analyzed. That said, the existence of this friction-like interaction is related to the quantization of vorticity and the interaction of said vortices with the excitation gas.

The two-fluid model was very successful at predicting *second sound* and *thermal counterflow*, and it is possible to connect, with certain restrictions, to the NLSE equation [22]. It is, however, not suitable to the study of complex vortex dynamics.

2.2.1.3 The vortex filament model

In a less microscopic approach, for the study of superfluid vortex dynamics it is not necessary to be concerned with the details of vortex nucleation, which can be investigated using the NLSE. A pure vortex dynamics model was developed by Schwarz [23], which relies on the interaction between the already existing vortices.

A vortex reacts to an external flow due to a *Magnus force*, which depends on the circulation about the vortex. If $\boldsymbol{s} = \boldsymbol{s}(\lambda, t)$ is a curve representing a vortex line at time t, and λ is the arc length, then the Magnus force is

$$\boldsymbol{F}_M = \kappa\rho_{\mathrm{s}}\boldsymbol{s}' \times (\boldsymbol{v}_\ell - \boldsymbol{v}_{T\mathrm{s}}), \tag{2.4}$$

where $\boldsymbol{v}_\ell = \frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}t}$ is the velocity of the vortex line and $\boldsymbol{v}_{T\mathrm{s}} = \boldsymbol{v}_{\mathrm{s}} + \boldsymbol{v}_i$ is the total superfluid velocity, consisting in the external imposed superfluid velocity and the self-induced velocity $\boldsymbol{v}_i$. The self induced velocity of the vortex depends on the its own curvature and is calculated through the Biot-Savart law,

$$\boldsymbol{v}_i(\boldsymbol{s}) = \frac{\kappa}{4\pi}\int \frac{(\boldsymbol{\zeta} - \boldsymbol{s}) \times \mathbf{d}\boldsymbol{\zeta}}{|\boldsymbol{\zeta} - \boldsymbol{s}|^3}, \tag{2.5}$$

where $\boldsymbol{s} = \boldsymbol{s}(\lambda, t)$ is a curve representing a vortex line at time t, and λ is the arc length, An approximation to this equation is usually used, known as the Local Induction Approximation (LIA),

$$\boldsymbol{v}_i(\boldsymbol{s}) \approx \beta\boldsymbol{s}' \times s'' \tag{2.6}$$

with $\beta = \kappa/(4\pi)\ln(1/|\boldsymbol{s}''|a_0)$, and a_0 is the vortex core radius. In the presence of a normal fluid component, a drag force, related to the mutual friction term in the two-fluid model, must be considered. Near the core of a vortex, this drag force can be written as

$$\boldsymbol{F}_d = -\frac{\kappa\rho_{\mathrm{s}}\rho_{\mathrm{n}}B}{2\rho}\boldsymbol{s}' \times (\boldsymbol{s}' \times (\boldsymbol{v}_{\mathrm{n}} - \boldsymbol{v}_{T\mathrm{s}})) - \frac{\kappa\rho_{\mathrm{s}}\rho_{\mathrm{n}}B'}{2\rho}\boldsymbol{s}' \times (\boldsymbol{v}_{\mathrm{n}} - \boldsymbol{v}_{T\mathrm{s}}), \tag{2.7}$$

where B and B' are the Hall and Vinen mutual friction coefficients. These coefficients are temperature dependent, and their values are known experimentally. Now, using the argument that the sum of forces on

the vortex line is zero, we obtain the Schwarz equation,

$$\boldsymbol{v}_\ell = \frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}t} = \boldsymbol{v}_\mathrm{s} + \boldsymbol{v}_i + \frac{\kappa\rho_\mathrm{s}\rho_\mathrm{n}B}{2\rho}\boldsymbol{s}' \times (\boldsymbol{v}_\mathrm{n} - \boldsymbol{v}_\mathrm{s} - \boldsymbol{v}_i) + \frac{\kappa\rho_\mathrm{s}\rho_\mathrm{n}B'}{2\rho}(\boldsymbol{v}_\mathrm{n} - \boldsymbol{v}_\mathrm{s} - \boldsymbol{v}_i), \tag{2.8}$$

This model has been very successful in simulations of complex vortex line geometries, with the cost of expensive computations if the Biot-Savart law is used, or loss of accuracy in the case of LIA. Furthermore, vortex reconnections must be handled explicitly in this model. The reconnection happens when the two vortices are sufficiently close together. This assumption, although justified through microscopic calculations, leads to an artificial component in the model. This dependence on the choice of model for the reconnection, although not too critical for a full Biot-Savart calculation, results in artifactual effects when using the LIA [24, 25], which might lead to misleading results.

2.2.1.4 The HVBK model

One way to model large amounts of vortices in a system is to consider a coarse-grained version of a superfluid. This model, known as the Hall-Vinen-Bekarevich-Khalatnikov (HVBK) model [20], is written as the two fluid model (eqs. 2.3) with additional terms to account for the effect of quantized vortices. The equations are

$$\frac{\mathrm{d}\boldsymbol{v}_\mathrm{s}}{\mathrm{d}t} = -\frac{1}{\rho}\nabla p - \frac{\rho_\mathrm{s}}{\rho_\mathrm{n}} S\nabla T + \nu_\mathrm{n}\nabla^2\boldsymbol{v}_\mathrm{n} + \boldsymbol{F} \tag{2.9a}$$

$$\frac{\mathrm{d}\boldsymbol{v}_\mathrm{n}}{\mathrm{d}t} = -\frac{1}{\rho}\nabla p + S\nabla T + \boldsymbol{T} - \frac{\rho_\mathrm{n}}{\rho}\boldsymbol{F}, \tag{2.9b}$$

with the definitions

$$\boldsymbol{\omega}_\mathrm{s} = \nabla \times \boldsymbol{v}_\mathrm{s} \tag{2.10a}$$

$$\hat{\boldsymbol{\omega}}_\mathrm{s} = \frac{\boldsymbol{\omega}_\mathrm{s}}{|\boldsymbol{\omega}_\mathrm{s}|} \tag{2.10b}$$

$$\boldsymbol{F} = \frac{B}{2}\hat{\boldsymbol{\omega}}_\mathrm{s} \times (\boldsymbol{\omega}_\mathrm{s} \times (\boldsymbol{v}_\mathrm{n} - \boldsymbol{v}_\mathrm{s} - \nu_\mathrm{s}\nabla \times \hat{\boldsymbol{\omega}}_\mathrm{s})) + \frac{B'}{2}\boldsymbol{\omega}_\mathrm{s} \times (\boldsymbol{v}_\mathrm{n} - \boldsymbol{v}_\mathrm{s} - \nu_\mathrm{s}\nabla \times \hat{\boldsymbol{\omega}}_\mathrm{s}) \tag{2.10c}$$

$$\boldsymbol{T} = -\nu_\mathrm{s}\boldsymbol{\omega}_\mathrm{s} \times (\nabla \times \hat{\boldsymbol{\omega}}_\mathrm{s}) \tag{2.10d}$$

$$\nu_\mathrm{s} = \frac{\kappa}{4\pi}\ln\left(\frac{b_0}{a_0}\right), \tag{2.10e}$$

where $\boldsymbol{F}$, $\boldsymbol{T}$, ν_s and b_0 are the friction, tension, vortex tension parameter and intervortex spacing respectively, with $b_0 = (2\omega_s/\kappa)^{-1/2}$, and B and B' are the Hall and Vinen coefficients. These equations assume that the vortex lines are somewhat spatially ordered, and are therefore not applicable to disordered tangles. The HVBK equations have been useful in describing the onset of complex superflows such as the Ekman to Taylor vortex flow transition [26], but they are too coarse grained for our purposes: the vortices are invisible, except for their collective effect.

2.2.2 The Ginzburg-Pitaevskii equations

Amongst all the successful models for superfluids, we have to single out a particular one, which is more microscopic model and is valid at temperatures close to the λ transition point. In that region, more attention must be paid to the normal fluid fraction if we are to consider a two-fluid model. This problem was tackled by Ginzburg and Pitaevskii [9, 10, 27] and they devised a very complete phenomenological model.

2.2.2.1 Equilibrium Condition

They proceeded the following way. To obtain the equations of motion for the superfluid we need to understand the equilibrium state the superfluid is relaxing to. To accomplish this we write its energy density

$$E = (\rho - |\psi|^2)\frac{v_n^2}{2} + \frac{1}{2}|\nabla\psi|^2 + \epsilon(\rho, S, |\psi|^2) \quad , \tag{2.11}$$

where ρ is the density of the superfluid, ψ is the complex scalar order parameter describing the pure superfluid part, $\boldsymbol{v}_n$ is the velocity of the normal component of the superfluid, and ϵ is the potential energy, of which we can specify the part dependent on the order parameter as the standard Landau-type energy, giving as a result

$$E = (\rho - |\psi|^2)\frac{v_n^2}{2} + |\psi|^2 - \frac{1}{2}\left(|\psi|^2\right)^2 + \frac{1}{2}|\nabla\psi|^2 + \epsilon_0(\rho, S). \tag{2.12}$$

To obtain the equilibrium condition this expression must be minimized with respect to $\boldsymbol{v}_n$ and ψ or ψ^*. But to do this properly we must also enforce the definition of the momentum of the superfluid,

$$\boldsymbol{j} = (\rho - |\psi|^2)\boldsymbol{v}_n + \rho_s\boldsymbol{v}_s \tag{2.13}$$

where, using the Madelung transformation $\psi = \rho_s^{1/2}e^{i\phi}$ and defining $\boldsymbol{v}_s = -\nabla\phi$ we can rewrite the momentum

expression as

$$\boldsymbol{j} = (\rho - |\psi|^2)\boldsymbol{v}_{\rm n} + \frac{i}{2}\left(\psi\nabla\psi^* - \psi^*\nabla\psi\right). \tag{2.14}$$

Finally, using a Lagrange multiplier $\boldsymbol{u}$ we include this definition into the expression for the energy density, yielding

$$E = (\rho - |\psi|^2)\frac{v_{\rm n}^2}{2} + |\psi|^2 - \frac{1}{2}\left(|\psi|^2\right)^2 + \frac{1}{2}|\nabla\psi|^2 + \boldsymbol{u}\cdot\left(\boldsymbol{j} - (\rho - |\psi|^2)\boldsymbol{v}_{\rm n} - \frac{i}{2}\left(\psi\nabla\psi^* - \psi^*\nabla\psi\right)\right) + \epsilon_0(\rho, S). \tag{2.15}$$

Now, minimizing with respect to ψ^* and $\boldsymbol{v}_{\rm n}$, eliminating $\boldsymbol{u}$ and assuming that the normal component of the superfluid is incompressible we arrive at the equilibrium condition

$$\frac{1}{2}\left(-i\nabla - \boldsymbol{v}_{\rm n}\right)^2\psi - (1 - |\psi|^2)\psi = 0 \tag{2.16}$$

2.2.2.2 Dynamical Equations

Given the expression for the energy we can expect ψ to obey a Schrödinger-like equation,

$$i\partial_t\psi = \hat{L}\psi \tag{2.17}$$

where $\hat{L}$ is a nonlinear operator. $\hat{L}$ cannot be purely Hermitian because the equation must account for the relaxation towards the equilibrium condition. We can readily specify the Hermitian part with the term

$$-\frac{1}{2}\nabla^2\psi + U\psi \tag{2.18}$$

where U is obtained from the expression for ϵ, giving

$$U = \left(-\frac{v_{\rm n}^2}{2} + 1 - |\psi|^2\right) - g \tag{2.19}$$

where g will be determined from conservation laws. The anti-Hermitian part drives the system towards the equilibrium condition specified in eq. 2.16. If the system is not too far from this equilibrium, then we can

write the anti-Hermitian (dissipative) part as

$$-i\Lambda\left\{\frac{1}{2}\left(-i\nabla-\boldsymbol{v}_{\mathrm{n}}\right)^{2}\psi-\left(1-|\psi|^{2}\right)\psi\right\}. \tag{2.20}$$

Λ is a parameter proportional to the inverse of the relaxation time and it must be a real number. Then, putting it all together, the equation of motion for the order parameter is

$$i\partial_t\psi=-\frac{1}{2}\nabla^2\psi-\left(1-\frac{1}{2}v_{\mathrm{n}}^2-|\psi|^2\right)\psi-g\psi-i\Lambda\left\{\frac{1}{2}\left(-i\nabla-\boldsymbol{v}_{\mathrm{n}}\right)^2\psi-(1-|\psi|^2)\psi\right\} \tag{2.21}$$

The mass conservation can be derived from the previous definitions, and is written

$$\partial_t\rho+\nabla\cdot\boldsymbol{j}=0. \tag{2.22}$$

The conservation of momentum is of the form

$$\partial_t j_i+\partial_k\Pi_{ik}=0 \tag{2.23}$$

where $\mathbf{\Pi}$ is the pressure tensor,

$$\Pi_{ik}=p\delta_{ik}+\rho_{\mathrm{n}}v_{\mathrm{n}i}v_{\mathrm{n}k}-\tau_{ik}+\frac{1}{4}\left\{\partial_i\psi\,\partial_k\psi^*-\psi^*\partial_i\partial_k\psi+\mathrm{c.c.}\right\}. \tag{2.24}$$

Here, p is the pressure, τ_{ik} is a symmetric tensor to be determined, and $\partial_i\equiv\frac{\partial}{\partial x_i}$. The equation for the entropy is

$$\partial_t S+\nabla\cdot\left(S\boldsymbol{v}_{\mathrm{n}}-\boldsymbol{q}\right)=\frac{R}{T}, \tag{2.25}$$

where $\boldsymbol{q}$ is the entropy flux, R is the dissipation function, and T is the temperature. Now, to get the still unknown coefficients g, τ_{ik}, $\boldsymbol{q}$ and R we must impose energy conservation in the form

$$\partial_t E+\nabla\cdot\boldsymbol{Q}=0 \tag{2.26}$$

where E is the energy density written in eq. 2.12. By expanding the $\partial_t E$ term, the unknown coefficients are

chosen so they cancel out the terms that cannot be written as a divergence. Those terms become

$$R = 2\Lambda\left|\left(\frac{1}{2}\left(-i\nabla-\boldsymbol{v}_{\rm n}\right)^2-\left(1-|\psi|^2\right)\right)\psi\right|^2+\frac{\kappa}{T}|\nabla T|^2+\frac{1}{2}\eta\left(\partial_k v_{{\rm n}i}+\partial_i v_{{\rm n}k}-\frac{2}{3}\delta_{ik}\nabla\cdot\boldsymbol{v}_{\rm n}\right)^2 \tag{2.27}$$

$$\boldsymbol{q} = \frac{\kappa}{T}\nabla T \tag{2.28}$$

$$\tau_{ik} = \eta\left(\partial_k v_{{\rm n}i}+\partial_i v_{{\rm n}k}-\frac{2}{3}\delta_{ik}\nabla\cdot\boldsymbol{v}_{\rm n}\right). \tag{2.29}$$

The term g and the part of τ_{ik} with non-zero trace have terms with the second viscosities not related to $\rho_{\rm s}$, and close to the λ point, where these equations are valid, the major contribution to second viscosities comes from the relaxation of $\rho_{\rm s}$. Thus g and the trace of τ_{ik} are set to zero. The constants κ and η are identified as the thermal conductivity and the first viscosity of the normal fluid, respectively. Finally we can write the system of equations that govern the dynamics of the superfluid close to the λ point:

$$i\partial_t\psi = -\frac{1}{2}\nabla^2\psi-\left(1-\frac{1}{2}v_{\rm n}^2-|\psi|^2\right)\psi-i\Lambda\left\{\frac{1}{2}\left(-i\nabla-\boldsymbol{v}_{\rm n}\right)^2\psi-\left(1-|\psi|^2\right)\psi\right\} \tag{2.30a}$$

$$\partial_t\rho+\nabla\cdot\boldsymbol{j} = 0 \tag{2.30b}$$

$$\partial_t j_i+\partial_k\Pi_{ik} = 0 \tag{2.30c}$$

$$\partial_t S+\nabla\cdot(S\boldsymbol{v}_{\rm n}) = \frac{1}{T}\nabla\cdot(\kappa\nabla T)+\frac{R}{T}, \tag{2.30d}$$

with the following definitions of the pressure tensor Π_{ik} and the entropy source function R,

$$\Pi_{ik} = p\delta_{ik}+\rho_{\rm n}v_{{\rm n}i}v_{{\rm n}k}-\eta\left(\partial_k v_{{\rm n}i}+\partial_i v_{{\rm n}k}-\frac{2}{3}\delta_{ik}\nabla\cdot\boldsymbol{v}_{\rm n}\right)+\frac{1}{4}\left\{\partial_i\psi\,\partial_k\psi^*-\psi^*\partial_i\partial_k\psi+\text{c.c.}\right\} \tag{2.31a}$$

$$R = 2\Lambda\left|\left(\frac{1}{2}\left(-i\nabla-\boldsymbol{v}_{\rm n}\right)^2-\left(1-|\psi|^2\right)\right)\psi\right|^2+\frac{\kappa}{T}|\nabla T|^2+\frac{1}{2}\eta\left(\partial_k v_{{\rm n}i}+\partial_i v_{{\rm n}k}-\frac{2}{3}\delta_{ik}\nabla\cdot\boldsymbol{v}_{\rm n}\right)^2, \tag{2.31b}$$

and $p=\frac{1}{2}\left(|\psi|^2\right)^2$ is the pressure. Together, equations 2.30 are known as the *Ginzburg-Pitaevskii Equations*. These equations are, in principle, only valid near the λ-point. But they can be seen as a generalized two-fluid model, as they are asymptotically correct in both the low- and high-temperature limits: the dissipation parameter $\Lambda\to 0$ as $T\to 0$, and $\Lambda\sim(T_\lambda-T)^{-1/3}$ for $T\to T_\lambda$.

We are particularly interested in this model because it provides a reasonable and straightforward mixture of two very similar formalisms, namely the NLSE –the Hermitian component– and the Ginzburg Landau

equation –the anti-Hermitian part. It is this line of thinking we are going to use to develop our model, although using a different modeling approach.

There is little record in the literature about the application of this equation to study superfluid problems. An interesting use is the numerical study of the spin-up problem in helium II made by Aranson & Steinberg [28], in which they only used equation 2.21 to characterize the nucleation and arrangement of vortices in a rotating container. There are a couple of studies by Aranson *et al.* [29] about vortex nucleation following a thermal quench [29, 30]. Also worth mentioning is the fact that the Ginzburg-Pitaevskii equations happen to be a special case of a very general framework for superfluid dynamics developed by Geurst [31].

To see how the order parameter equation 2.21 relates to Landau and Tisza's two-fluid model (eq. 2.3, it is useful to perform a Madelung transformation $\psi = \rho_s^{1/2} e^{i\phi}$ with $\boldsymbol{v}_s = \nabla\phi$. After some calculation, equation 2.21 now reads

$$\partial_t \rho_s + \nabla \cdot (\rho_s \boldsymbol{v}_s) = \Lambda \left\{ \rho_s^{1/2} \nabla^2 \rho_s^{1/2} - (\boldsymbol{v}_s - \boldsymbol{v}_n)^2 \rho_s + 2(1-\rho_s)\rho_s \right\} \tag{2.32a}$$

$$\partial_t \boldsymbol{v}_s = \nabla \left[\rho_s^{-1/2} \nabla^2 \rho_s^{1/2} - \frac{1}{2}\left(v_s^2 + v_n^2\right) + (1-\rho_s) + \frac{\Lambda}{2\rho_s} \nabla \cdot (\rho_s(\boldsymbol{v}_s - \boldsymbol{v}_n)) \right]. \tag{2.32b}$$

It is evident from the above equations that in the dissipation-less limit ($\Lambda \to 0$) they become the equations for a NLSE fluid. Also it is interesting to see that the counterflow velocity term $\boldsymbol{v}_s - \boldsymbol{v}_n$ only appears in the dissipative terms (containing the factor Λ).

2.2.2.3 Remarks on the Ginzburg-Pitaevskii equations

The Ginzburg-Pitaevskii equations are strongly nonlinear, and hence their behavior will differ from the relatively simple NLSE dynamics. A useful check is to see how the motion of a 2D vortex is modified by these nonlinearities. For the case $v_n \ll 1$, and treating the nonlinearities in eq. 2.21 in a perturbative manner, it is shown in Pismen [1] that the Magnus force is modified by the dissipative terms. (The equivalent of equation 2.30a in Pismen [1] differs only by the simple scaling transformation $\nabla \longrightarrow \nabla/\sqrt{2}$, $\boldsymbol{v}_n \longrightarrow \boldsymbol{v}_n/\sqrt{2}$.)

$$\boldsymbol{v}_s = \boldsymbol{v} + \frac{1}{1+\Lambda^2}\left[-\Lambda^2(\boldsymbol{v} - \boldsymbol{v}_n) + \mathcal{J}(\boldsymbol{v} - \boldsymbol{v}_n) N \Lambda \ln \frac{v_0(1+\Lambda^2)}{\Lambda(\boldsymbol{v} - \boldsymbol{v}_n)} \right], \tag{2.33}$$

where $\boldsymbol{v}$ is the vortex velocity, v_0 is a constant dependent on the vortex core structure, N is the winding number and $\mathcal{J}$ is a fully anti-symmetric unit matrix,

$$\mathcal{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{2.34}$$

In this form, for $|N| = 1$, v_0 takes the value $v_0 \approx 3.29$. Compare this expression to equation 2.38 in Carlson [32], which rewritten to our variables reads

$$\boldsymbol{v}_\mathrm{s} = \boldsymbol{v} + N\alpha\zeta\mathcal{J}(\boldsymbol{v} - \boldsymbol{v}_\mathrm{n}), \tag{2.35}$$

where $\mathbf{K}$ is the phase gradient, $\dot{\mathbf{Q}}$ is the vortex velocity, n is the winding number, α is a constant dependent on core structure and ζ is the dissipation coefficient. Note that equations 2.33 and 2.35 both contain a similar set of constants, but eq. 2.33 contains a logarithmic nonlinearity. This nonlinearity comes from the asymptotic matching with the far region, on which, for $\Lambda \neq 0$, the phase obeys a Poisson equation instead of a Laplace equation.

For illustrative purposes, we will rewrite equation 2.30a in amplitude (polar) form. By making the substitution $\psi = Ue^{i\phi}$ and defining $\boldsymbol{v}_\mathrm{s} = \nabla\phi$ we get

$$\partial_t U = -\boldsymbol{v}_\mathrm{s}\cdot\nabla U - \frac{1}{2}U\nabla\cdot\boldsymbol{v}_\mathrm{s} + \Lambda\left(\frac{1}{2}\nabla^2 U - \frac{1}{2}U(\boldsymbol{v}_\mathrm{s} - \boldsymbol{v}_\mathrm{n})^2 + \left(1 - U^2\right)U\right) \tag{2.36a}$$

$$\partial_t \boldsymbol{v}_\mathrm{s} = \nabla\left[\frac{1}{2U}\nabla^2 U - \frac{1}{2}\left(v_\mathrm{s}^2 + v_\mathrm{n}^2\right) + \left(1 - U^2\right) + \frac{1}{2U}\Lambda\left((\boldsymbol{v}_\mathrm{s} - \boldsymbol{v}_\mathrm{n})\cdot\nabla U + \nabla\cdot(U(\boldsymbol{v}_\mathrm{s} - \boldsymbol{v}_\mathrm{n}))\right)\right] \tag{2.36b}$$

And if we identify U as $U^2 = \rho_\mathrm{s}$ we can write the equations 2.36 in more physically meaningful notation:

$$\partial_t\rho_\mathrm{s} + \nabla\cdot(\rho_\mathrm{s}\boldsymbol{v}_\mathrm{s}) = \Lambda\left\{\rho_\mathrm{s}^{1/2}\nabla^2\rho_\mathrm{s}^{1/2} - (\boldsymbol{v}_\mathrm{s} - \boldsymbol{v}_\mathrm{n})^2\rho_\mathrm{s} + 2(1 - \rho_\mathrm{s})\rho_\mathrm{s}\right\} \tag{2.37a}$$

$$\partial_t \boldsymbol{v}_\mathrm{s} = \nabla\left[\rho_\mathrm{s}^{-1/2}\nabla^2\rho_\mathrm{s}^{1/2} - \frac{1}{2}\left(v_\mathrm{s}^2 + v_\mathrm{n}^2\right) + (1 - \rho_\mathrm{s}) + \right.$$
$$\left. + \frac{\Lambda}{2\rho_\mathrm{s}}\nabla\cdot(\rho_\mathrm{s}(\boldsymbol{v}_\mathrm{s} - \boldsymbol{v}_\mathrm{n}))\right] \tag{2.37b}$$

First, it can be rearranged to

$$\partial_t \psi + i2\Lambda \boldsymbol{v}_\mathrm{n} \cdot \nabla \psi = (i + \Lambda)\left[\frac{1}{2}\nabla^2\psi + \left(1 - \frac{1}{2}v_\mathrm{n}^2 - |\psi|^2\right)\psi\right]. \tag{2.38}$$

From here we make the substitution $\psi = Ue^{i\phi}$, and we then get the system

$$\begin{aligned}
\partial_t U - 2\Lambda U \boldsymbol{v}_\mathrm{n} \cdot \nabla\phi &= \Lambda\left[\frac{1}{2}\nabla^2 U + \left(1 - \frac{1}{2}(\boldsymbol{v}_\mathrm{n}^2 + |\nabla\phi|^2) - U^2\right)U\right] \\
&\quad - \left[\frac{1}{2}U\nabla^2\phi + \nabla U \cdot \nabla\phi\right] \qquad (2.39a) \\
U\left(\partial_t\phi + 2\Lambda \boldsymbol{v}_\mathrm{n} \cdot \nabla U\right) &= \Lambda\left[\frac{1}{2}U\nabla^2\phi + \nabla U \cdot \nabla\phi\right] \\
&\quad - \left[\frac{1}{2}\nabla^2 U + \left(1 - \frac{1}{2}(\boldsymbol{v}_\mathrm{n}^2 + |\nabla\phi|^2) - U^2\right)U\right]. \qquad (2.39b)
\end{aligned}$$

As a final remark, we must note that equation 2.30a is equation 8.12 in Geurst [31] (with appropriate dimensions), where he also gives the corresponding kinetic coefficients (the matrix $\boldsymbol{\alpha}$) for the generalized complex Ginzburg-Pitaevskii equations.

2.3 Cell dynamical systems

We now turn to the method that we use to implement a sensible model for superfluids that will try to include all the relevant physics.

We are interested in the long-time behavior of superfluid systems, where macroscopic features dominate the dynamics. Any tool that enables this kind of study must account for the symmetries, boundary conditions and any essential features, while at the same time safely disregarding irrelevant microscopic dynamics.

To obtain a computationally efficient description we use cell-dynamical systems [33] (CDS) to simulate the dynamics of the order parameter. A cell-dynamical system is a map from a set of discrete patterns to itself. This definition also applies to other discrete systems such as cellular automata and coupled map lattices. As an example, a finite difference map is a CDS, but not all CDS are finite difference discretizations. In this sense, CDS gives us greater freedom in choosing a model and the relevant variables for the problem to study. Another benefit of CDS modeling is that there is no great need for tuning of the system, as long as the parameters describe a plausible physical situation.

CDS maps can be created in many ways. For instance, to model a Ginzburg–Landau equation, we note that it has a Laplacian and a potential term. The Laplacian simply couples each cell with its neighbors, therefore it can be implemented by subtracting from each cell value the mean around it, being careful to be as isotropic as possible. In this way gradients are penalized. The potential term simply drives the system towards an equilibrium value. In the CDS way of thinking, any map that has the correct fixed points for the potential is in the same universality class as the physical problem, so the final choice depends on the kind of control needed on the map, and also the numerical efficiency must be considered. Another example, and the first success of the CDS approach, is a conservative spinodal decomposition CDS map [33], whose dynamics are governed by the Laplacian of a functional derivative of the free energy, has nested averaging and potential maps. And in a final example, the CDS maps used to model travertine terraces found in hot springs [34] are just very simple rules governing the fluid dynamics and physicochemical dynamics of the sedimentation process.

The CDS approach uses the following reasoning. One can model nature by differential equations, and then discretize them to get a set of coupled maps. Or one can devise the coupled map description directly from modeling nature. Both types of models extract from experience the key dynamics of the underlying processes, but they express them in different ways. In principle, they should be equivalent, since they are both abstractions and expressions of the same physics. This argument was quantitatively tested in the example of a complex model for carbonate flow depositional patterns, as found in the vicinity of geothermal hot springs. In this study, Goldenfeld *et al.* [8] solved a cylindrically-symmetric flow deposition problem obtaining the shape of a travertine (calcium carbonate) dome by two methods. The first was a CDS method, the second was a direct integration of the fluid dynamics equations for shallow-water flow coupled to the moving boundary problem representing the growing calcium carbonate surface. The resulting patterns were identical. Moreover, the partial differential equations yielded an analytical scaling prediction for a particular morphological characteristic, and this scaling law was shown to be satisfied over 5 decades by the CDS numerical solution. Thus, two logically independent representations of the same underlying physics were demonstrated to yield the same predictions.

It may appear that CDS is a rather crude approach, but the truth is that we are considering *only* the relevant dynamical processes needed to calculate universal scaling properties, and that has been enough to deliver testable predictions. As an example, Nagaya *et al.* [7] experimentally measured vortex correlation

functions in 2D liquid crystal films, and the universal scaling agreed with the CDS predictions with no adjustable parameters. See Zapotocky *et al.* [35] for another example involving defect dynamics liquid crystals, where the order parameter is not scalar, but matricial. CDS maps have been used to study spinodal decomposition [36], coupled oscillators [37] and vortex string dynamics [5, 6]. It has also been successfully used with more tangible results to understand ordering kinetics in liquid crystals [35] and to explain the formation of travertine terraces in hot springs [34]. We choose to use a CDS map for our problem because it allows us to capture the symmetries of the superfluid in a concise way, and also, since it gives us latitude to choose any map satisfying the correct physics, it allows for very fast numerical simulation.

2.4 A CDS-based fast computational algorithm for superfluids

To construct the CDS map for our model we first must note the structure present in equation 2.21. By construction, it has a conservative, Hermitian part and a dissipative, anti-Hermitian part. If we were only to consider the dynamics of the Hermitian part, we would be dealing with the Gross-Pitaevskii equation with a slightly modified chemical potential term. In an analogous way, the behavior of the anti-Hermitian part follows the Ginzburg-Landau equation, with an extra term representing the advection of the order parameter field due to the presence of the normal fluid with velocity $\boldsymbol{v}_{\mathrm{n}}$.

For the first, conservative part we note that it is a set of coupled oscillators with an amplitude-dependent frequency. The oscillator frequency is, in principle, dependent on the magnitude of the order parameter and the normal component of the superfluid. Thus, this map will have the form

$$\psi^{n+1} = \exp\left(-i\omega(|\psi|^2, \boldsymbol{v}_{\mathrm{n}})\right)\psi^n. \tag{2.40}$$

This map is also the solution semigroup for the Gross-Pitaevskii equation, as usually obtained in CDS models. By inspecting eq. 2.21, the frequency term $\omega(|\psi|^2, \boldsymbol{v}_{\mathrm{n}})$ can be directly written as

$$\omega(|\psi|^2, \boldsymbol{v}_{\mathrm{n}}) = 1 - \frac{1}{2}v_{\mathrm{n}}^2 - |\psi|^2 \tag{2.41}$$

To perform the spatial coupling, we note that equation 2.21 has only Laplacian terms, so it is sufficient

to penalize gradients. Thus, the CDS map for the Hermitian operator is

$$\psi^{n+1} = \exp\left(-i\left(1 - \frac{1}{2}v_{\mathrm{n}}^2 - |\psi|^2\right)\right)\psi^n + iC\left(\langle\psi^n\rangle - \psi^n\right) \tag{2.42}$$

where C sets the spatial coupling strength, and $\langle\ldots\rangle$ denotes an isotropic spatial average.

Following the same reasoning, we turn to the dissipative part of equation 2.21. As this part includes an advection term, it is useful to treat it separately. For this map, the complex order parameter here is no longer viewed as an oscillator, but its magnitude is driven towards the superfluid fixed point. The CDS map is thus written then as

$$\psi^{n+1} = f(A, \psi^n, \boldsymbol{v}_{\mathrm{n}}) + C\left(\langle\psi^n\rangle - \psi^n\right) \tag{2.43}$$

The function f can be any function that satisfies the flow set by the double-well potential used in equation (2.21). Explicitly, given a value of the parameter A, the fixed point at $\psi = 0$ is repulsive, together with the attractive region $|\psi| = 1$, which represents the superfluid state or, for a different set of values for A, only a single, attractive fixed point at $\psi = 0$. Note that the normal velocity $\boldsymbol{v}_{\mathrm{n}}$ enters in this part too, but to keep the map simple we just write this term expliticly. Our choice for the function f is

$$f(A, \psi) = \frac{A\psi}{\sqrt{1 + |\psi|^2(A^2 - 1)}}. \tag{2.44}$$

For $A < 1$ the fixed point at $\psi = 0$ is attractive, and repulsive for $A > 1$. This parameter sets the depth of the quench, how fast the system evolves towards its fixed point. A practical reason to use this map is the fact that the function $1/\sqrt{x}$ is implemented in hardware in modern CPUs, allowing for noticeably faster computation (avoiding a division).

Putting everything together at this point, the algorithm works as follows:

$$\psi^{n+1/3} = \lambda\left\{\exp\left(-i\left(1 - \frac{1}{2}v_{\mathrm{n}}^2 - |\psi|^2\right)\right)\psi^n + iC\left(\langle\psi^n\rangle - \psi^n\right)\right\} \tag{2.45a}$$

$$\psi^{n+2/3} = \mathrm{Advection}\left(\psi^{n+1/2}, \boldsymbol{v}_{\mathrm{n}}\right) \tag{2.45b}$$

$$\psi^{n+1} = \frac{A\psi^{n+2/3}}{\sqrt{1 + |\psi^{n+2/3}|^2(A^2 - 1)}} + \frac{1}{2}v_{\mathrm{n}}^2 + C\left(\left\langle\psi^{n+2/3}\right\rangle - \psi^{n+2/3}\right) \tag{2.45c}$$

where $\lambda \in [0, 1]$ is a parameter quantifying the relative strength of the conservative and dissipative parts, occupying a role similar to $1/\Lambda$, where Λ is the dissipation parameter that appears in eq. 2.21.

Now, the average $\langle \ldots \rangle$ needs to be more isotropic [38] than the regular first order Laplacian traditionally used in finite-difference calculations. The averages used are [6, 36]

$$\langle \psi \rangle \equiv \frac{1}{6} \sum_{\text{NN}} \psi + \frac{1}{12} \sum_{\text{NNN}} \psi \qquad \text{for 2D} \tag{2.46}$$

$$\langle \psi \rangle \equiv \frac{1}{9} \sum_{\text{NN}} \psi + \frac{1}{36} \sum_{\text{NNN}} \psi \qquad \text{for 3D,} \tag{2.47}$$

where "NN" means the nearest-neighbor cells, ans "NNN" means the next to nearest-neighbor cells.

The advection algorithm is, in our case, a simple scalar product of the first derivative of the order parameter with the normal velocity. Care must be placed, nonetheless, in the choice of numerical implementation of this first derivative. Again, demanding a high degree of isotropy [38] and accurate representation in Fourier space [39], we choose the first derivative as

$$\partial_x \psi \equiv \frac{1}{8} \left(\psi_{i+1,j+1} + 2\psi_{i,j+1} - \psi_{i-1,j+1} + \psi_{i+1,j-1} - 2\psi_{i,j-1} - \psi_{i-1,j-1} \right) \tag{2.48}$$

for two dimensions, and

$$\begin{aligned} \partial_x \psi &= \frac{1}{8} \left(\psi_{i+1,j+1,k} - \psi_{i-1,j+1,k} + \psi_{i+1,j-1,k} - \psi_{i-1,j-1,k} + \right. \\ &+ \left. \psi_{i+1,j,k+1} - \psi_{i-1,j,k+1} + \psi_{i+1,j,k-1} - \psi_{i-1,j,k-1} \right) + \\ &+ \frac{1}{4} \left(\psi_{i,j+1,k} - \psi_{i,j-1,k} + \psi_{i,j,k+1} - \psi_{i,j,k-1} \right), \end{aligned} \tag{2.49}$$

for three dimensions, with the appropriate swapping of indices to obtain the other components of the gradient. Failure to use a highly isotropic derivative will result in artifacts in the simulation.

2.4.1 Remarks on the model

This model has many interesting features. First, vortices are emergent and not explicitly put in the model and, as such, will naturally undergo reconnection. In presence of a normal flow, the vortices will move according to a Magnus force.

The model, thanks to its cellular description, permits studying large numerical systems and should reach the asymptotic regime faster that conventional realizations of superfluid simulations [40]. The model does not need to care about microscopic details: the CDS maps are coarse-grained descriptions of the fluids.

Although it was inspired by the Ginzburg-Pitaevskii equations, they are not a discretization for it. In the same spirit as other cellular models, it is meant to be *per se* a model of simple quantum fluids.

So far the model has not explicitly included thermodynamical terms similar to the ones in the Ginzburg-Pitaevskii equations. This means that we are operating in a regime not too far away from the λ transition point. Also, there are no temperature gradients (and hence no thermal counterflow is possible). See Aranson & Steinberg [28] for details. In principle, even when there are no external temperature gradients, there exist flutuations due to the friction between the normal and superfluid parts. This also means that in the approximation we are using, the normal fluid velocity should be much smaller than the speed of first sound. It is possible to account for these thermodynamical effects by coupling the model of eqs. 2.45 to a reasonable fluid model such as Lattice Boltzmann. More on this on Sec. 2.5. Another limitation is that the model does not reveal the microscopic dynamics of the superfluid, but the model does not try to do that in the first place. For most of our calculations, this will not be a problem.

2.5 Lattice Boltzmann

In the model written in eqs. 2.45 we do not explicitly account for the dynamics of the normal fluid, or even less the thermodynamics of it. So far, this normal velocity $\boldsymbol{v}_{\mathrm{n}}$ is an external forcing field, much like an external magnetic field acting on a superconductor. We can couple the CDS model to a comparatively simple model for normal fluids through $\boldsymbol{v}_{\mathrm{n}}$, as long as we can affect the normal fluid dynamics with the superfluid order parameter. Here the task is apparently more difficult as we must dwell into the well-established realm of computational fluid dynamics (CFD). CFD models have been very successful in simulating fluids, even in multiple scales, but these successes are often backed by massive amounts of computational power. In our case we are not interested in fluids at high or even moderately high Reynolds numbers, but this does not mean we are discarding the interesting physics of complex superflows. One method that satisfies these criteria, and which happens to be a very successful one too, is the Lattice Boltzmann equation.

The Lattice Boltzmann equation [13, 14] (LBE) is a discrete counterpart to the Boltzmann equation. It can be expressed in a very simple form as

$$f_\alpha(\boldsymbol{x} + \boldsymbol{e}_\alpha, t + 1) - f_\alpha(\boldsymbol{x}, t) = \Omega_\alpha \,. \tag{2.50}$$

Here, $f(\boldsymbol{x}, t)$ is a particle distribution function, which contains the particles which, at time t and lattice site $\boldsymbol{x}$ are moving along the discrete velocity vectors $\boldsymbol{e}_\alpha$, and Ω_α is the collisional operator. As with the case of the continuous Boltzmann equation, the moments of the distribution function are related to macroscopic physical quantities of the fluid:

$$\rho = \sum_\alpha f_\alpha \qquad \boldsymbol{j} = \rho \boldsymbol{u} = \sum_\alpha \boldsymbol{e}_\alpha f_\alpha \,. \tag{2.51}$$

To use the LBE first one must make assumptions about the collisional operator. The simplest nontrivial operator assumes that the particle distribution f relaxes to an equilibrium distribution f^{eq} with a single relaxation time. That is,

$$\Omega_\alpha = -\frac{1}{\tau}\left(f_\alpha - f_\alpha^{\mathrm{eq}}\right) . \tag{2.52}$$

This is known as the Bhatnagar-Gross-Krook operator. Despite its apparent oversimplification, it captures enough information which enables the system to simulate realistic fluid dynamics. The key lies in the modeling of f^{eq}. In most numerical studies, f^{eq} is a simple polynomial on the fluid velocity and lattice velocity vectors, whose coefficients are set through the reasonable requirement that the moments of f^{eq} are the same as those of f. The coefficients are not uniquely set, and they depend on the geometry of the lattice being used.

There is previous work in the area of complex fluid models being simulated using LB methods. These studies include phase separation in Cahn-Hilliard hydrodynamics and defect dynamics in liquid crystal systems [41]. Most of the methods for the inclusion of the complex fluid dynamics were taken from these studies. They used the pressure tensor Π_{ik} as an extra restriction on the equilibrium distribution, and only then the correct fluid dynamics were recovered.

Taking the hint from those studies, we can model f^{eq} by requiring

$$\sum_\alpha f_\alpha^{\mathrm{eq}} e_{\alpha i} e_{\alpha k} = \Pi_{ik}. \tag{2.53}$$

In the classical case, the fluid stress tensor is explicitly dependent on the fluid velocity, that is, with terms linear in $v_i v_k$, so the expansion for f^{eq} must at least include terms up to second order in velocities.

In the Lattice Boltzmann formalism, to make the connection with hydrodynamic equations we must perform a Chapman-Enskog expansion (valid for low Mach numbers) together with an appropriate choice of the collision operator Ω, just like in the case for the Boltzmann equation. The choice we make (and probably

the simplest choice) is the Bhatnagar-Gross-Krook operator. This operator drives the system toward an equilibrium distribution specified by f^{eq} with a relaxation time τ. Our task is to appropriately model f^{eq} to match our requirements.

Given the following equations for the fluid,

$$\partial_t \rho + \nabla \cdot \boldsymbol{j} = 0 \tag{2.54a}$$

$$\partial_t j_i + \partial_k \Pi_{ik} = 0 \tag{2.54b}$$

one way to model f^{eq} is to enforce that its moments satisfy the relations

$$\sum_\alpha f_\alpha^{\mathrm{eq}} = \rho \tag{2.55a}$$

$$\sum_\alpha f_\alpha^{\mathrm{eq}} \boldsymbol{e}_\alpha = \boldsymbol{j} \tag{2.55b}$$

$$\sum_\alpha f_\alpha^{\mathrm{eq}} e_{\alpha i} e_{\alpha k} = \Pi_{ik}\,. \tag{2.55c}$$

This is true as long as Π_{ik} is symmetric. If it's not, then only the symmetric part enters into this equation, and the anti-symmetric part will enter as a body force via an additional set of constraints.

2.5.1 The Chapman-Enskog expansion

As a check before using Lattice Boltzmann methods, I have to make sure the prescription above is actually simulating the conservation equations. For this I perform a Chapman-Enskog expansion on equation 2.50. This expansion takes the continuum limit of 2.50 and, after matching order by order, it is possible to obtain the first few terms of the particle distribution function f as a function of the equilibrium distribution f^{eq}. From there, it is straightforward to obtain the evolution equations of the moments of f, up to the prescribed order.

To proceed with the expansion, first I Taylor-expand the left-hand side of the Lattice Boltzmann equation:

$$f_\alpha^{\mathrm{out}} - f_\alpha = f_\alpha + e_{\alpha i}\partial_i f_\alpha + \frac{1}{2} e_{\alpha i} e_{\alpha j} \partial_i \partial_j f_\alpha + \partial_t f_\alpha + \partial_t e_{\alpha i} \partial_i f_\alpha + \frac{1}{2}\partial_t^2 f_\alpha - f_\alpha\,, \tag{2.56}$$

where $f_\alpha^{\text{out}} \equiv f_\alpha(\boldsymbol{x} + \boldsymbol{e}_\alpha, t + 1)$. We thus obtain the following expression for the LBGK equation:

$$e_{\alpha i}\partial_i f_\alpha + \frac{1}{2}e_{\alpha i}e_{\alpha j}\partial_i\partial_j f_\alpha + \partial_t f_\alpha + \partial_t e_{\alpha i}\partial_i f_\alpha + \frac{1}{2}\partial_t^2 f_\alpha = -\frac{1}{\tau}\left(f_\alpha - f_\alpha^{\text{eq}}\right) \tag{2.57}$$

Now, I expand the distribution function up to order 2,

$$f_\alpha = f_\alpha^{(0)} + f_\alpha^{(1)} + f_\alpha^{(2)} + \ldots \tag{2.58}$$

and together with the assumption that the derivatives will increase by one the order of a term in the expansion, it is found that

$$f_\alpha^{(0)} = f_\alpha^{\text{eq}} \tag{2.59a}$$

$$f_\alpha^{(1)} = -\tau\left(e_{\alpha i} + \partial_i\right) f_\alpha^{\text{eq}} \tag{2.59b}$$

$$f_\alpha^{(2)} = \tau(1 - \frac{1}{2\tau})\left(e_{\alpha i} + \partial_i\right)^2 f_\alpha^{\text{eq}}. \tag{2.59c}$$

We can now find the evolution equations for the moments. The relaxation time τ can take values greater or equal than $\frac{1}{2}$. The case $\tau = \frac{1}{2}$ is the "non-viscous" limit, as it does not introduce terms into the evolution equations resembling the viscous terms in the Navier-Stokes equation. These viscous terms, with careful tuning of the value of τ, are usually an advantage in most Lattice Boltzmann simulations. However, given the fields I am simulating in this calculation, I chose to include the correct viscous terms in the definition of Π_{ik} and not to include the Lattice Boltzmann viscous terms up to second order in f_α.

2.5.1.1 Non-viscous limit

After setting the value of τ to $\frac{1}{2}$, the right-hand side of equation (2.59c) vanishes, leaving us with the following terms in equation 2.58, the expansion of f,

$$f_\alpha = f_\alpha^{\text{eq}} - \frac{1}{2}\left(e_{\alpha i} + \partial_i\right) f_\alpha^{\text{eq}} + \text{h.o.t.} \tag{2.60}$$

Now, summing the above equation over α, using the definitions 2.51 and the constraints 2.55 we obtain

$$\partial_t \rho + \partial_i j_i = 0 + \text{h.o.t.}\ , \tag{2.61}$$

that is, we recovered the continuity equation. To get the next equation we multiply eq. 2.60 by $e_{\alpha q}$ and again summing over α, and using the definitions and constraints, we find

$$\partial_t j_i + \partial_k \Pi_{ik} = 0 + \text{h.o.t.} \,, \tag{2.62}$$

thus recovering the fluid equations I need to simulate.

2.5.1.2 Viscous limit

To show why I use the non-viscous LB prescription, and for the sake of completeness, I repeat the derivation of the equations, this time with $\tau > \frac{1}{2}$.

The expansion of f has the form

$$f_\alpha = f_\alpha^{\text{eq}} - \tau \left(e_{\alpha i} + \partial_i\right) f_\alpha^{\text{eq}} + \tau(1 - \frac{1}{2\tau}) \left(e_{\alpha i} + \partial_i\right)^2 f_\alpha^{\text{eq}}. \tag{2.63}$$

By doing the same manipulations as before we obtain the following:

$$\partial_t \rho + \partial_i j_i = \left(\tau - \frac{1}{2}\right) \{\partial_t \left[\partial_t \rho + \partial_i j_i\right] + \partial_i \left[\partial_t j_i + \partial_k \Pi_{ik}\right]\} \tag{2.64a}$$

$$\partial_t j_k + \partial_i \Pi_{ik} = \left(\tau - \frac{1}{2}\right) \left\{\partial_t \left[\partial_t j_i + \partial_i \Pi_{ik}\right] + \partial_i \left[\partial_t \Pi_{ik} + \partial_q \mathbb{M}_{ikq}\right]\right\}, \tag{2.64b}$$

where $\mathbb{M}_{ikq} = \sum_\alpha f_\alpha^{\text{eq}} e_{\alpha i} e_{\alpha k} e_{\alpha q}$. In equation 2.64a we note that the first term in brackets is the same as the right hand side, so it can be discarded based on the high order of that term. The same can be said for equation 2.64b. Thus, to first order in derivatives, we have the following equations:

$$\partial_t \rho + \partial_i j_i = 0 \tag{2.65a}$$

$$\partial_t j_k + \partial_i \Pi_{ik} = 0. \tag{2.65b}$$

Now, replacing 2.65b into the second term in brackets in equation 2.64a we can get rid of a yet higher order

term. Thus, up to second order in derivatives, we get the final set of equations

$$\partial_t \rho + \partial_i j_i = 0 + O(\partial^3) \tag{2.66a}$$

$$\partial_t j_k + \partial_i \Pi_{ik} = \left(\tau - \frac{1}{2}\right) \partial_i \left[\partial_t \Pi_{ik} + \partial_q \mathbb{M}_{ikq}\right] + O(\partial^3). \tag{2.66b}$$

Now, from the definition of f^{eq} and using identities for the lattice vectors, we can calculate $\partial_q \mathbb{M}_{ikq}$:

$$\partial_q \mathbb{M}_{ikq} = \frac{1}{3}\left(\partial_k j_i + \partial_i j_k + \partial_q j_q \delta_{ik}\right) \tag{2.67}$$

Finally, defining $\eta = \frac{1}{3}\left(\tau - \frac{1}{2}\right)$ we obtain the final evolution equations for the moments of f up to second order in derivatives,

$$\partial_t \rho + \partial_i j_i = 0 \tag{2.68a}$$

$$\partial_t j_k + \partial_i \Pi_{ik} = \partial_i \left\{\eta \left[\partial_k j_i + \partial_i j_k + \partial_q j_q \delta_{ik}\right] + 3\eta \partial_t \Pi_{ik}\right\} \tag{2.68b}$$

It is easy to see from these equations the effect of the single relaxation time $\tau > 1/2$: it creates the viscous stress-like terms. This is useful for Navier-Stokes simulation of other types of fluids. Instead, given we are simulating complex fluids, it is more appropriate to include the appropriate viscous terms in the definition of Π_{ik}, and set $\tau = 1/2$.

2.5.2 Coupling back to the order parameter equations: backreaction

Up to this point we have stated a model for the order parameter (in eqs. 2.45), which has a normal fluid term forcing it, and a model for this normal fluid, written as a Lattice Boltzmann description. Now we need to close the model by coupling this LB model back to the order parameter description.

To do this, we note that in treating complex fluids using LB, the pressure tensor Π_{ik} must be specified explicitly. In other systems, the pressure tensor is defined naturally through a free energy [41]. In our case, we look back at the Ginzburg-Pitaevskii equations in 2.21. We recall that, although this model does not come from a free energy, it does have an explicit term for the pressure tensor that arises from conservation

of momentum, which is written in eq, 2.31a. It reads

$$\Pi_{ik} = \frac{1}{2}\left(|\psi|^2\right)^2 \delta_{ik} + \rho_{\rm n} v_{{\rm n}i} v_{{\rm n}k} - \eta\left(\partial_k v_{{\rm n}i} + \partial_i v_{{\rm n}k} - \frac{2}{3}\delta_{ik}\nabla\cdot\boldsymbol{v}_{\rm n}\right) + \frac{1}{4}\left\{\partial_i\psi\,\partial_k\psi^* - \psi^*\partial_i\partial_k\psi + \text{c.c.}\right\} \tag{2.69}$$

In principle we can calculate this quantity and use it in the constraint equations for the LB model. There are many things to note. First, the viscosity η in eq. 2.69 is not the same as the one inferred in the Chapman-Enskog expansion, and it must be specified as another parameter of the model. Second, care must be taken when justifying the Chapman-Enskog expansion when comparing the momentum term $\boldsymbol{j}$ from a regular fluid description with the rather complex momentum in the Ginzburg-Pitaevskii equations.

Despite these caveats, it is a promising avenue worth exploring. There are thermodynamics implicit in the Lattice Boltzmann description and, for example, a temperature gradient can be imposed as another contraint equation.

2.6 Validation of the models

Presently, only the passive model in eqs. 2.45 are implemented, that is the cellular maps with the advection routine and no backreaction. With this it is possible to test the model to see how well it reproduces known features of superfluid hydrodynamics, namely reconnections and their reaction to the presence of a normal velocity field. Also we can validate how well the model reproduce known results for defects in Ginzburg-Landau systems.

2.6.1 Two- and three-dimensional quenches

The first most crucial feature that the model must satisfy is that it actually supports vortex solutions. We already know that in a pure cellular XY model [5, 6], topological defects are created as a result of a critical quench. What is left to see is what happens after adding the coupled oscillator map and the advection routine.

In principle, the coupled oscillator map should allow for defects, as it was modeled after the NLSE. The effect of the advection routine is more difficult to analyze on its own, as it depends on the actual form of the external velocity field, as it will be seen in the section on kinematic simulations. Nevertheless, a simple test case of uniform, constant velocity suffices to see if we were in the right track.

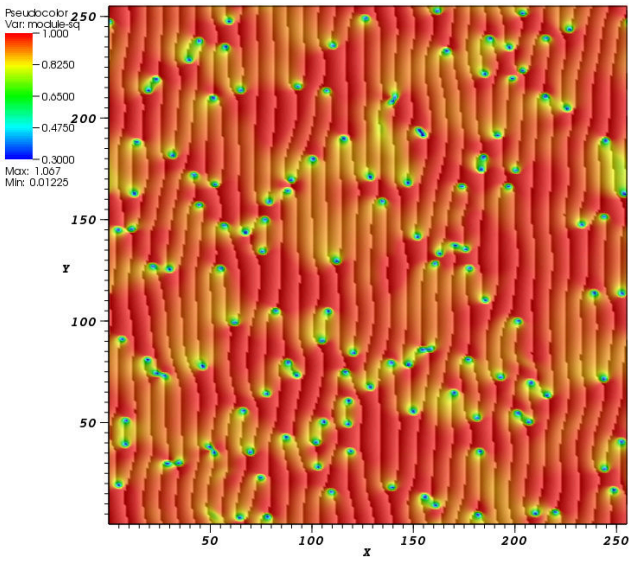


FIGURE 2.1: Two-dimensional quench with external velocity field. The lightness indicates the phase and the color scale indicates the density of the superfluid. The abrupt changes in lightness accounts for a 2π phase slip.

By looking at the equations of motion of the vortices, the external velocity field should manifest itself as a externally imposed phase gradient, so a simple uniform field will look like a constant phase gradient, modulo 2π along the direction of the velocity field. Also, the Magnus force should be evident as vortices annihilate (in the 2D case) or reconnect (in the 3D case). To simulate the quench, we follow the recipe outlined in Mondello & Goldenfeld [5, 6]. The simulations were initialized using random complex numbers, whose phase $\phi \in [0, 2\pi)$ is uniformly distributed, and whose magnitude is $|\psi| = 1 \pm 0.1$, also uniformly distributed. That way we are starting close to the superfluid state but with a disordered initial condition. The externally imposed velocity field is set to be $\boldsymbol{v}_{\mathrm{n}} = 0.5\hat{\boldsymbol{x}}$.

In two dimensions the quench created a set of topological defects, which can be identified as the amplitude zeros of the field. Also, the defects are quantized vortices: the flow circulates about them and there is a phase discontinuity. In Figure 2.1, the density of the field $|\psi|^2$ is plotted, with the color scale going from minimum (blue) to maximum (red). The superimposed light/dark shade shows the phase field, with the sharp transition from light to dark represents a 2π jump in the phase value. The defects are always located at an extreme of a phase discontinuity, and about every vortex there is only one jump, indicating that they are singly-quantized vortices. We cannot rule out the possibility of the existence of higher order singularities, but in a simple superfluid these are unstable.

The vortices do annihilate as they come close together, and the actual annihilation dynamics vary de-

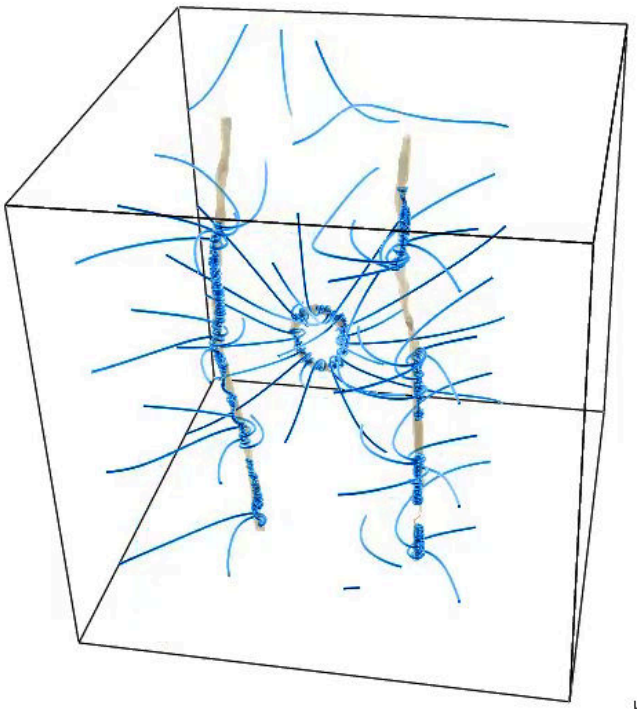


FIGURE 2.2: Superfluid streamlines wrap around two line vortices and a ring vortex.

pending on the presence or not of the external field. Also, this simple external field has the expected effect of creating an overall background of a constant phase gradient along the horizontal direction. From the image in Figure 2.1 we can make our case for the analogy that the defects in superfluids are similar objects to dislocations in a crystal lattice.

In three dimensions, the quench results in the creation of a tangle of one-dimensional defects, which, like their two-dimensional siblings, are transported by the external field. These singularities undergo reconnection upon contact and once they form rings (see Figure 2.2), they shrink until disappearing.

2.6.2 Vortex reconnection scaling

One of the important features of superfluid hydrodynamics is the fact that, upon contact, quantum vortices reconnect, and when doing so they dissipate some energy by locally giving off phonons, creating a rarefaction wave [42]. In simulations based on vortex string models, in which the vortices are the only objects being simulated, they interact through the Biot–Savart Law or the Local Induction Approximation [23] , but the reconnection event must be artificially added, with the consequence that this relevant dissipation mechanism is neglected, and these results might be misleading. Also in the case of the LIA, the non-local

effects are not accounted for, so the asymptotic dynamics of vortex tangles are not correct in that framework [24].

In the case of our model, there is no need to add any artificial rule to implement reconnections: they come naturally with our prescription. Given that they are topological defects, do they behave as they should? This class of defects, which also include one-dimensional cosmic strings [1], has a very definite scaling law: the distance between two reconnecting strings scales as $(t_0 - t)^{1/2}$ or $(t - t_0)^{1/2}$, depending if the strings are closing in together or moving away from each other, respectively. In superfluid helium II, it was confirmed experimentally that the superfluid vortices do obey this scaling law [43] although the distribution of calculated exponents raised questions about the effect of the local medium in the neighborhood of the event. Also, some corrections to this law might be needed due to intermittency effects [44]. It is a good test to see if our model obeys this scaling law.

2.6.2.1 Measurement procedure and results

The following procedure was used to measure the exponent for the scale parameter. First we identify a reconnection event, its position and approximate time. Then we isolate this region of interest from the rest of the string tangle to aid in the visualization (see the isolated reconnection event snapshots in Fig. 2.3). To measure the distance we have to note that immediately after the reconnection event a plane can be defined such that the reconnected strings will pass through this plane perpendicularly. The segment joining the points defining the distance between the strings (the shortest possible distance between the strings) should lie in this plane. With the visualization software we make a 2D slice of the density data, placing it approximately where this ideal plane would be located (small deviations from the ideal position of the plane are not relevant because the coefficients multiplying the position of the string cores will affect both of them equally, thus the distance scaling will remain unaffected). The density data in this slice shows the location of the string cores and therefore the distance is easily computed. A similar analysis can be made for the instant previous to the reconnection.

As the pre-reconnection scale parameter obeys the power law $\ell(t) = a(t_0 - t)^{\beta}$, and the post-reconnection scale parameter obeys $\ell(t) = a(t - t_0)^{\beta}$, with β having a predicted value of $1/2$,

We have to do an extra bit of work to deal with this scaling law's dependence on the reconnection time t_0. Instead of performing multiple fits for the exponent using various values for t_0 and then choosing the

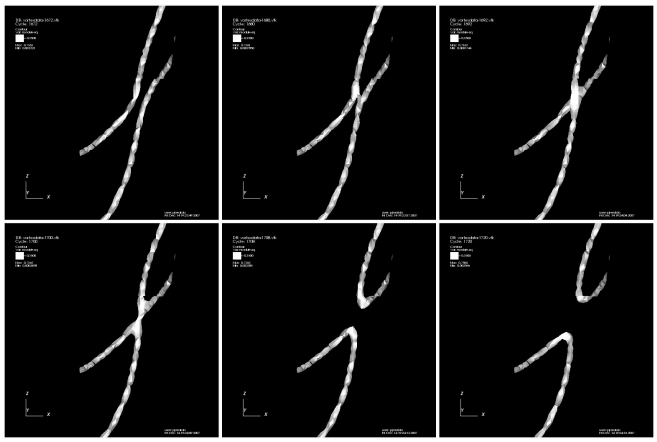


FIGURE 2.3: Frames extracted from the simulation showing the evolution of the reconnection event.

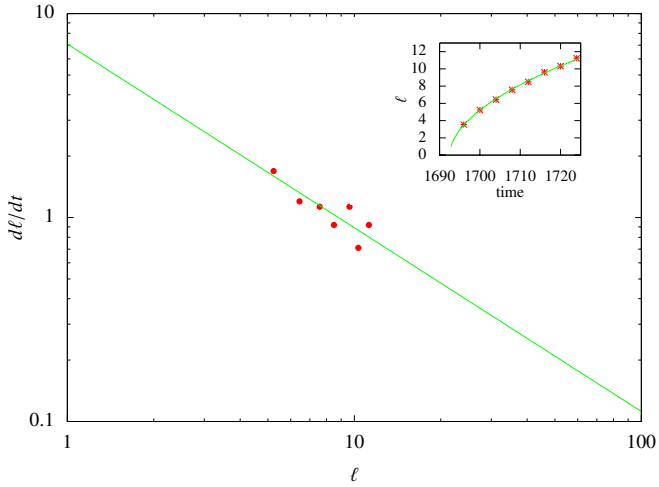


FIGURE 2.4: Scaling for two reconnecting vortex strings. Inset: distance as a function of time, with its best power-law fit.

best one, we proceed the following way [45]: we calculate the time derivative of $\ell(t)$, and we plot it as a function of $\ell(t)$. If $\ell(t)$ has a power law dependence as we expect, so will $\ell'(\ell(t))$. We only have to relate the exponents. Assuming $\frac{d\ell}{dt} = \ell^{\alpha}$, and $\ell(t) = a(t - t_0)^{\beta}$, then it is straightforward to get the leading exponent for $\ell(t)$, $\ell \propto t^{1/(1-\alpha)}$, finally obtaining $\beta = 1/(1 - \alpha)$.

From our current data we find the values listed in Table 2.1. In particular, we find $\alpha = -0.9 \pm 0.2$, thus obtaining $\beta = 0.53 \pm 0.06$, in good agreement with the prediction. Using these values we obtained $t_0 = 1652.4 \pm 0.2$ and $a = 1.78 \pm 0.01$ (see Fig. 2.4, inset). Note that although these quantities have nonstandard units, these values can be useful for making comparisons of adimensionalized quantities.

Although this measuring method is cumbersome, and thus the data is scarece, these results indicate that the topological defects displayed by our model seem to obey the same scaling laws as superfluid vortices.

$\alpha \pm \delta\alpha$	$\beta \pm \delta\beta$	type
-0.9 ± 0.2	0.53 ± 0.06	post
-1.0 ± 0.4	0.5 ± 0.1	pre
-1.0 ± 0.2	0.50 ± 0.05	pre
-0.9 ± 0.6	0.5 ± 0.2	post

TABLE 2.1: Values obtained for the scaling exponents for different reconnection events. For the reconnection type column, *pre* means that the scaling parameter was measured immediately before the reconnection. Likewise, *post* means that the measurement was performed immediately after the reconnection.

Later we will discuss some other much stronger scaling results that have the consequence that the exponent β indeed has a value close to 0.5 for our CDS model.

2.6.3 Kinematic calculations

A series of studies were performed using the model, to replicate known results and verify that some mechanisms present in superflows are captured by the model. These particular simulations were kinematic in nature, in the sense that an external normal flow is imposed, and then the superfluid fraction response was observed.

2.6.3.1 Ostermeier-Glaberson Instability and the Arnold–Beltrami–Childress Flow

When the normal velocity component which is parallel to a vortex line exceeds a critical value, an instability of the Kelvin waves takes place. This phenomenon is known as the Ostermeier-Glaberson (OG) instability [46], and is responsible for the creation of vortex lines by using energy extracted from the normal fluid flow.

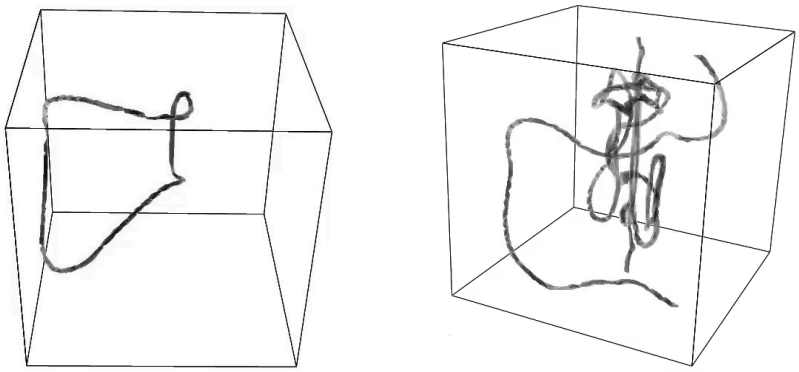
To test if our model can reproduce this phenomenon, we advect the superfluid with a Gaussian vortex flow [47]:

$$\boldsymbol{v}_{\mathrm{n}}(\boldsymbol{r}) = \frac{\Gamma}{2\pi r}\left(1 - \exp(-r^2/r_c^2)\right)\hat{\boldsymbol{\phi}} \tag{2.70}$$

This flow displays concentrated vorticity at its core, the expected behavior is that the vortex lines will get trapped and stretched, increasing the line density.

The simulations were carried in a periodic cubic lattice with 64^3 cells, and the vortex centered in the cube, with its axis parallel to the $\hat{z}$ direction. The initial condition was a single ring inside the cube, a remainder of a quench run, and various values for the circulation parameter Γ were used in different runs, ranging from $\Gamma = 4$ to $\Gamma = 12$.

For a small value of Γ tried, there was already some behavior which corresponds to the sought after



(a) Onset of the OG instability. Initial twisting before the creation of the first ring.

(b) A more developed tangle about the Gaussian vortex

FIGURE 2.5: Different stages of the Ostermeier-Glaberson instability.

instability. The initial ring is trapped by the vortex. It is then stretched around the vortex core and twisted. (see Fig. 2.5(a)). A ring is created once the quantum vortex lines reconnect, which also becomes transported by the normal flow. This ring reconnects again with its parent, creating Kelvin waves. For this small Γ, the normal flow is large enough to repeat the process. But for larger values, the Kelvin waves are rapidly stretched by the normal flow, creating more rings as they become twisted. From this we see that the vortex line length increases exponentially, as it is expected for this instability. These simulations also showed that the vortex line length eventually saturated. This was predicted in Samuels [47], but the author could not reach that regime.

One of the most important tests that applied to this model is advecting the superfluid flow with a normal flow complex enough that it will be capable to induce chaotic, yet not turbulent, behavior of the vortex strings, to test some statistical features that should be expected in any reasonable model for superfluids. This particular flow is known as the Arnold–Beltrami–Childress (ABC) flow [48, 49].

$$v_x = C_y \cos\left(\frac{2\pi}{\lambda} y\right) + C_z \sin\left(\frac{2\pi}{\lambda} z\right) \tag{2.71}$$

$$v_y = C_z \cos\left(\frac{2\pi}{\lambda} z\right) + C_x \sin\left(\frac{2\pi}{\lambda} x\right) \tag{2.72}$$

$$v_z = C_x \cos\left(\frac{2\pi}{\lambda} x\right) + C_y \sin\left(\frac{2\pi}{\lambda} y\right) \tag{2.73}$$

where the C_k are the flow's parameters.

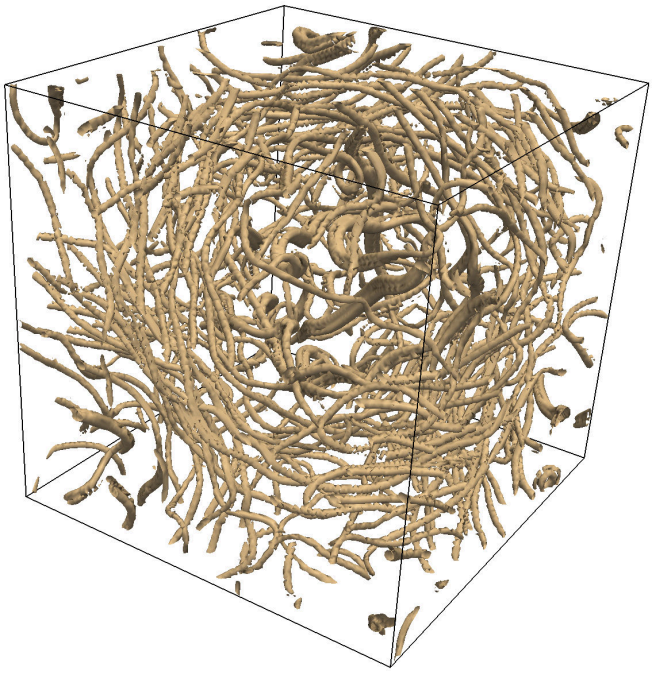


FIGURE 2.6: Snapshot of vortex tangle under the ABC flow. Although the tangle is a complex object, it is definitely not complex enough to say that it is a turbulent state.

The ABC flow has a number of desirable features. It displays staggered tubes of concentrated vorticity, with strongly interacting fluid lines, allowing to explore, for some parameters, chaotic Lagrangian trajectories [48]. It also has non-zero helicity, $\boldsymbol{h} = \boldsymbol{v} \cdot \boldsymbol{\omega}$. It is also a simple analytic solution of the NSE, which allows for comparison of exact results with the data from simulations.

The simulations aim to reproduce two features. Namely, from a simple initial condition the line density should increase exponentially until saturation, due to the Ostermeier-Glaberson instability. Also, the mean vorticity of the superfluid should match the vorticity of the advecting ABC field.

The following results assume that the parameters C_k are identical, $C_k = C$. A simulation was performed on a periodic lattice with 128^3 cells. The initial condition was a pair of opposing vortex rings, described by its Clebsch potentials [50]

$$\lambda(\boldsymbol{x}) = \cos\left(\frac{2\sqrt{2}}{L}|y - y_0|\right) \qquad \mu(\boldsymbol{x}) = \sin\left(\frac{8\pi}{3L}|\boldsymbol{x} - \boldsymbol{x}_0|\right), \tag{2.74}$$

and then, after a brief relaxation time, the system is left evolving under different values of C for 20000 iterations. Six values of C were explored, from 0.025 to 0.15 in steps of 0.025.

The RMS vorticity scales linearly with the line density in a fully turbulent regime [51]. As seen in

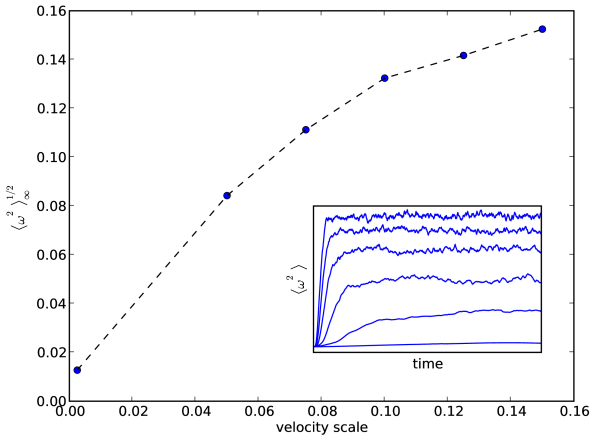


FIGURE 2.7: RMS vorticity as a function of input velocity scale. In a turbulent state this dependence should be linear. Inset: saturation of enstrophy as a function of time. The saturation level increases as the forcing of the normal fluid increases. Note how the fluctuations also increase.

the in figure 2.7, we don't see that linear dependence, probably indicating we are not in a fully developed turbulent state. Inset in the same figure, the mean enstrophy $\langle\omega^2\rangle$ increases exponentially for all velocities, except $C = 0.025$, which is likely below the threshold for the OG instability. Also noticeable is the increase in fluctuations. But probably more remarkable is the saturation of the vorticity in itself. In the study by Barenghi *et al.* [49], they cannot obtain the saturation of vorticity using the Local Induction Approximation, as it discards the non-local effects needed. The saturation was expected because, in the framework of the Biot–Savart law (BSL) , as the superfluid velocity approaches the normal velocity, the mutual friction, which drives the superflow in the first place, diminishes. They also could not use the BSL directly as it is computationally impractical. Their solution was to derive an *ad hoc* approximation by assuming that the superflow was close in geometry to the normal flow (which is known analytically), and subsequently informing this approximation by sparsely sampling the simulation field using the BSL. Only then the authors could obtain the expected saturation. Although this approximation matches the BSL in the first stages, where the LIA already lost validity, it requires that the normal advecting flow has a known functional form, making it unsuitable for quasi-turbulent or turbulent flows. Our model includes non-local effects and vortex reconnection in a natural way, and has the advantage that the computational effort scales only with system size, not with vortex density. This allows for prolonged computation in complex, dense vortex tangles.

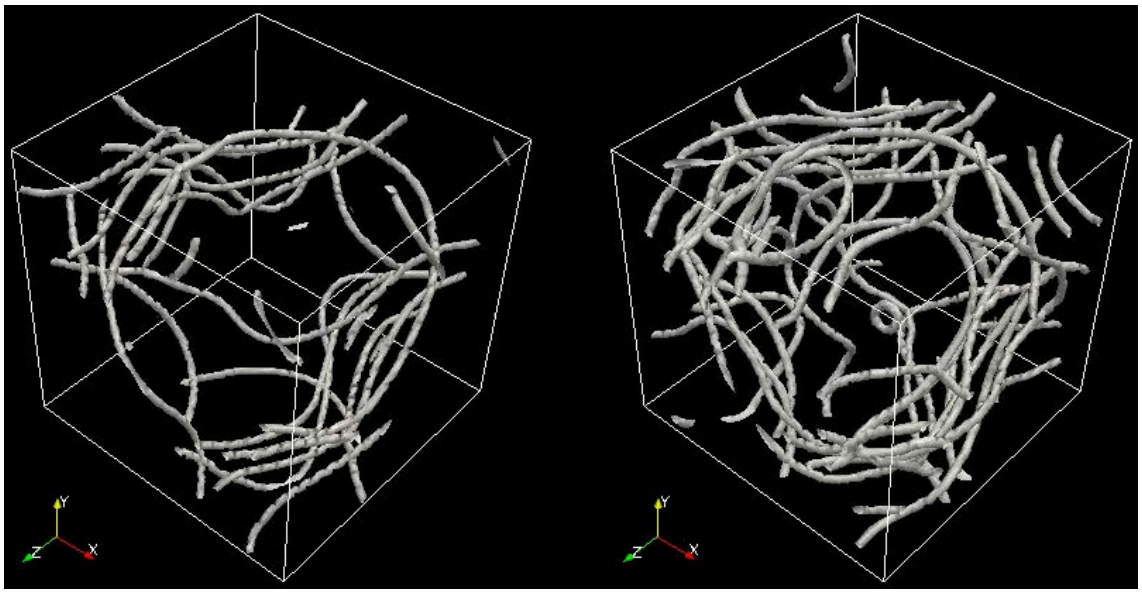


FIGURE 2.8: Comparison between two different realizations of the CDS model in three dimensions forced with ABC flow. In both situations, the external imposed flow has the same parameters. In the left frame, the CDS model is fully dissipative ($\lambda = 0$ in eq. 2.45a), whereas on the right $\lambda = 0.25$. Comparing both situations, having a conservative component makes the system more prone to the creation of the defect strings.

2.6.3.2 Comparison between fully dissipative and mixed models in ABC flow

To end this section, a natural question to ask is the extent of the effect of the conservative terms in the CDS maps written in eqs. 2.45. The external flow adds an extra phase contribution to the order parameter, and also the Magnus force between vortices changes direction as the value of Λ changes, as seen in eq. 2.33. Thus, it would be expected both an increase in the number of vortices formed (due to the extra phase term) and changes in the motion of the vortices in response to their interaction.

To explore these effects, I set up a simulation of a three-dimensional system subject to an ABC flow. One of the simulations was fully dissipative, with $\lambda = 0$ in eq. 2.45a, and one with some conservative effect, with $\lambda = 0.25$, keeping everything else identical. As seen in Fig. 2.8, both resulting fields show a similar structure of the vortex tangle, following the vortex tube structure of the ABC flow. However, the simulation on the right (with the conservative maps turned on) displays a higher density of strings, confirming that this fluid is indeed more prone to creation of defects. Otherwise, both dynamics seem very similar, with intermittent rearrangements of the vortex tangle due to the chaotic streamlines of the ABC flow.

2.6.4 Spin-up of superfluids

A very particular experiment that can be performed on superfluids is the spin-up of the fluid [52]. The basic setup of the problem consists in subjecting the superfluid to rotation, resulting in the formation of vortices in its bulk. The big difference, compared to regular fluids, is that due to the quantization of circulation, a lattice of vortices will be created, much like an Abrikosov lattice in superconductors, instead of a single vortex representative of all the circulation of the fluid.

In the framework of numerical models, Aranson & Steinberg [28] used the Ginzburg-Pitaevskii model of superfluids [9, 10, 27] to study the spin-up problem close to the superfluid transition. The setup of the problem is very simple and easy to implement with the CDS model.

Instead of creating a circular boundary condition to represent the container, the authors used simple periodic boundary conditions. The container was represented by adding an attenuating factor to the Ginzburg-Pitaevskii equation of the form

$$\partial_t \Psi = d(r)\Psi, \tag{2.75}$$

where the attenuating function $d(r)$ has the form

$$d(r) = \begin{cases} 0 & d < R \\ -\sinh(0.5\sqrt{r^2 - R^2}) & d \geq R, \end{cases} \tag{2.76}$$

where r is the distance to the center of the container, and R is the radius of the container. The effect of this attenuation is that the superfluid inside is unaffected, there is attenuation at the wall (with a defined penetration depth) and outside the container there is no superfluid at all.

Driving this superfluid is an external normal flow resulting from a rigidly rotating solution of the normal fluid equations in the Ginzburg-Pitaevskii system, where the azimuthal velocity is described by

$$\boldsymbol{v}_{\mathrm{n}} = \Omega r \hat{\phi} \tag{2.77}$$

where Ω is the angular velocity of the rotating fluid.

I implemented this prescription for the problem in the CDS model, and compared the results with and without the conservative term.

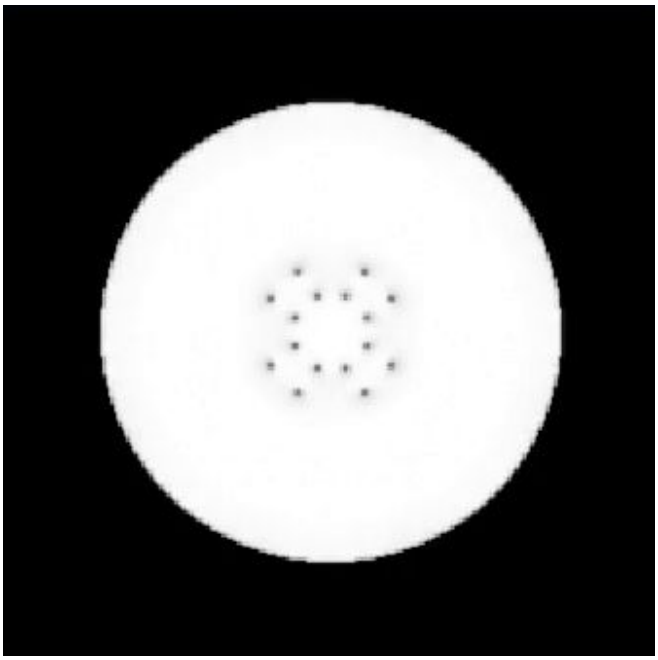


FIGURE 2.9: Abrikosov-like lattice of quantized vortices inside a superfluid described by dissipative dynamics. The lattice is very ordered, where vortices created from the wall move towards the center, then repelling each other as they form the lattice and reach a stationary regime.

In the first experiment, the CDS model was run with the conservative maps turned off, that is $\lambda = 0$ in eq. 2.45a. As seen in Fig. 2.9, a lattice of vortices is indeed created. This lattice has crystal-like conformation, which is stable once is reaches equilibrium. The vortices are created at the walls and move towards the center, and as they get closer to each other, the effective repulsion between them becomes evident, creating the resulting lattice. As expected the number of vortices increases as the value of Ω is increased. On an existing lattice, if one slightly increases the velocity, another vortex is created and migrates towards the lattice, and all vortices eventually rearrange to reach a new stable configuration. Compare to Fig. 3c in Aranson & Steinberg [28].

In the second experiment, the conservative part of the CDS map was turned on, setting $\lambda = 0.25$. As seen in Fig. 2.10, an arrangement of vortices is still formed, but it never reaches an ordered configuration. Instead, vortices do not remain stationary and move around the container, avoiding each other and exhibit density waves emanating from them as they move. Also, for the same value of Ω, the fluid is more prone to create defects in the system.

To conclude, this simple model of spin-up exhibits the expected features of superfluids under rotation. Simple extensions are possible, including cylindrical and Taylor-Couette geometries. Far from being a toy

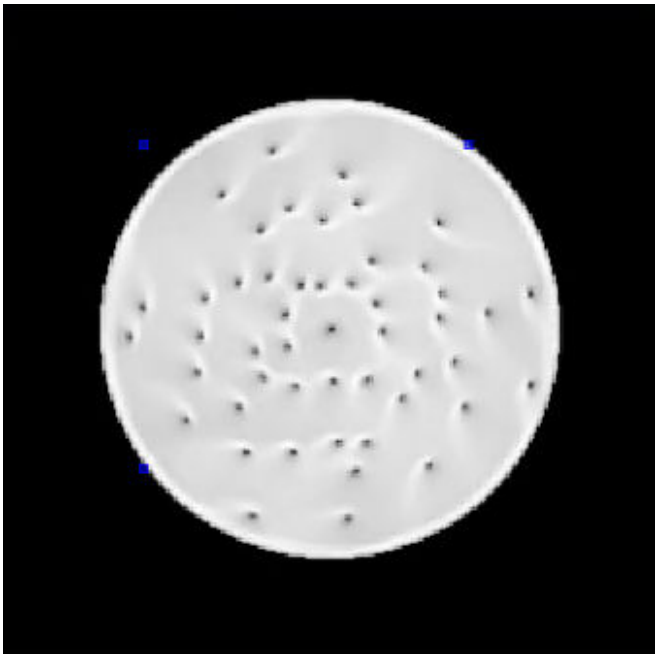


FIGURE 2.10: Vortex arrangement at the center of the superfluid, modeled with both a conservative and dissipative components. The system does not reach an ordered state, and the vortices remain unsettled and move around the container. With the same forcing angular velocity, the system is more prone to create defects

model, a simple spin-up system can give useful information about more complex geometries. For example, a model of neutron stars is a rotating sphere of superfluid matter [52], and one of the unexplained features of them, namely pulsar glitches, has been modeled with a very simple spin-up system similar to the one described above, with an added external potential for pinning of vortices [53], exhibiting vortex arrangement avalanches due to depinning of vortices, which is thought to cause these glitches.

2.6.5 Scaling results

CDS models have been very successful at modeling and predicting universal behavior, such as exponents for asymptotic scalings. To validate our model, I show that our system exhibits scaling regimes consistent with what we expect for a quantum fluid system.

In particular, when a system with $O(n)$ symmetry undergoes a critical quench, its structure factor obeys a scaling regime known as a generalized Porod's law (see Bray & Humayun [54] and references therein). This law states that, during a critical quench, the high-k tail of the structure factor scales [54] as

$$S(k,t) \propto \rho_{\mathrm{def}}(t) k^{-(d+n)}, \tag{2.78}$$

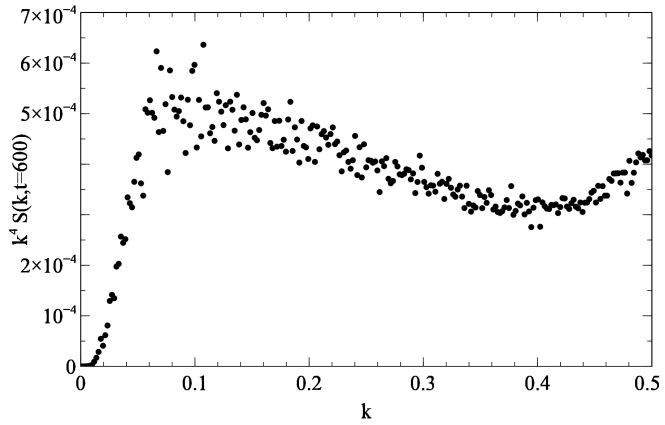


FIGURE 2.11: Scaling plot of the structure factor for an $O(2)$ system in two dimensions, expecting a k^{-4} scaling. A straight line parallel to the x axis denotes an exact exponent of -4. Note the range of the y axis.

where ρ_{def} is the density of defects, and d is the dimensionality of the system. The particular form of the structure factor is written as

$$S(k,t) = \langle \phi_k(t)\phi_{-k}(t)\rangle, \tag{2.79}$$

where $\phi_k(t)$ is the Fourier transform of the order parameter $\phi(t)$, and $\langle \ldots \rangle$ is an average over realizations of the quench.

It is of note that in eq. 2.78 there are two predictions about the behavior. One, the exponent of the tail. And two, the density term as it evolves over time.

Testing these predictions is relatively straightforward. We setup a critical quench simulations, as in Mondello & Goldenfeld [5], and calculate the structure factor following eq. 2.79. Since our system has $O(2)$ symmetry and we implement it in two dimensions, the expected tail should behave as k^{-4}.

In Fig. 2.11 we do a scaling plot of $k^{-4}S(k,t)$ as a function of k, to test if the structure factor behaves as expected. Since the curve only slightly deviates from the horizontal (note the range of the y axis), we can claim that our CDS system behaves as expected for an $O(2)$ system.

Now, to test for the scaling of the defect density, we setup two different simulations. One is the quench just described, and the other one is a quench while the system is sheared with an external velocity field. Under shear, the number of defects should reach a stationary value, as opposed to a quench situation where all defects will eventually get annihilated.

In Fig. 2.12 we plot the structure factor value, evaluated at the last calculated value of k, as a function of time. The quantity is plotted for both the quench and the sheared system, and is expected to be proportional

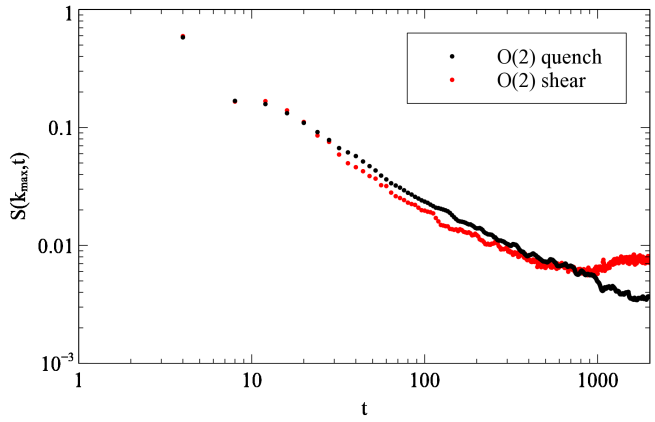


FIGURE 2.12: Plot of the structure factor evaluated at the maximum calculated value of k, as a function of time for both a quench and a sheared system. This value is expected to be proportional to the defect density in the system. In both cases the system dramatically reduce the density of defects as they are annihilated. For the sheared system, this quantity reaches a minimum and then equilibrates, as opposed to the quenched system where in continues to decrease.

to the density of defects in the system. In both cases this quantity decreases as a power law, as is expected for the density of defects in a quench. For the sheared system, the quantity eventually reaches a minimum, then equilibrates to a stationary value. For the case of the quench, the quantity keeps decreasing until it eventually reaches the limit of numerical accuracy. We can conclude then that the behavior of this quantity is consistent with it being proportional to the density of defects.

Another possible scaling law that can be verified and its related to Porod's law, is the scaling of the velocities of defects as they undergo a quench. We detail this particular study in Chapter 3.

2.6.6 Tracking of defect velocities

As part of the research performed, we were interested in the velocity statistics of the topological defects, such as superfluid vortices and crystal dislocations. Since the defects we are interested in can be found by means of the zeros of the order parameter field, we can, in principle, simply numerically locate the zeros and follow them as the simulation evolves (see Qian & Mazenko [55] for a relevant example). Although the premise is simple and does give us the desired quantities, it also involves significant overhead, which would mean that proper statistics gathered from these quantities would take a long time to obtain. Moreover, extending this idea to three dimensions, where defects are strings instead of points, makes this an all but impossible task to accomplish within a reasonable time frame.

Let's look at the question more carefully. On one hand, since we are looking for velocity statistics, we

realize that using the above method we get more information than needed, such as position information for each defect. (There are interesting situations where such information is desired, such as tracking of the lifetime of defect pairs; see Huepe *et al.* [56] for an example.) On the other hand, our simulations give us access to the full order parameter field, whose dynamics have encoded a wealth of information about the system. The question then becomes: "is it possible to draw information of the velocity of the topological directly from the order parameter field?" A secondary question then is what is the price we have to pay in efficiency for this information.

Fortunately for us, this mapping between order parameter and defect velocity exists. It was originally mentioned by Halperin [11], and then extensively used and expanded by Mazenko [12] in analytical research of topological defects [12, 57–59]. It uses the fact that these topological defects are defined by the zeros of the order parameter, and also that the topological charge of the system obeys a conservation law. What we obtain from this is the velocity vector, in the case of a point defect, and the velocity vector of a line element in the case of a defect string. We can also extract the topological charge of the defect, including its sign (and charge density for defect strings). Curiously, we don't know of any numerical implementation of this method prior to our own work.

2.6.6.1 Derivation of the method

The method works as follows. Let our order parameter field be represented by $\boldsymbol{\psi}(\boldsymbol{x})$. This order parameter has N components, and the coordinates live in a space with D dimensions. We are mostly interested in point defects ($N = D$) and defect strings ($N = D - 1$).

We want to go from the field variables to the particle (defect) variables. The natural way to do this is to consider the determinant of the Jacobian [12]

$$\mathcal{D} = \left\| \frac{\partial \psi_\alpha}{\partial x_\beta} \right\| \tag{2.80}$$

where $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_N$ and $\beta = \beta_1, \beta_2, \ldots, \beta_D$ are the indices for the field and coordinate components, respectively. In particular, we can write, for the $N = D$ case of point defects, the topological charge density as

$$\rho(\boldsymbol{x}) = \delta(\boldsymbol{\psi})\mathcal{D}(\boldsymbol{x}) \tag{2.81}$$

where δ is the Dirac Delta distribution, and the Jacobian is written as

$$\mathcal{D} = \frac{1}{N!}\epsilon_{\beta_1\beta_2\ldots\beta_N}\epsilon_{\alpha_1\alpha_2\ldots\alpha_N}\nabla_{\beta_1}\psi_{\alpha_1}\nabla_{\beta_2}\psi_{\alpha_2}\ldots\nabla_{\beta_N}\psi_{\alpha_N} \tag{2.82}$$

where ϵ is a fully antisymmetric symbol. Then we write the fact that the topological charge is conserved, using a continuity equation,

$$\partial_t\rho + \nabla_\beta(\rho v_\beta) = 0 \tag{2.83}$$

where $\boldsymbol{v}$ is the defect velocity field, and is written as

$$v_\beta = \frac{-1}{\mathcal{D}}\frac{1}{(N-1)!}\epsilon_{\beta\beta_2\ldots\beta_N}\epsilon_{\alpha_1\alpha_2\ldots\alpha_N}\dot{\psi}_{\alpha_1}\nabla_{\beta_2}\psi_{\alpha_2}\ldots\nabla_{\beta_N}\psi_{\alpha_N} \tag{2.84}$$

where $\dot{\psi}_{\alpha_1}$ is the time derivative of ψ_{α_1}. We are interested in the statistics of the velocity field $\boldsymbol{v}$.

For the sake of clarity, let's derive the formula for $\boldsymbol{v}$ in the case for point defects in two dimensions and a complex scalar order parameter. The scalar order parameter can be written as (see also Chapter 3)

$$\psi = \psi_1 + i\psi_2 \tag{2.85}$$

and the topological charge density defined in eq. 2.81. In our two dimensional system the topological charge is conserved and obeys equation 2.83. In particular, the determinant of the Jacobian of the transformation reads [60]

$$\mathcal{D} = \frac{1}{2i}\left(\nabla_x\psi^*\nabla_y\psi - \text{c.c.}\right) \tag{2.86}$$

where "c.c." stands for "complex conjugate". By differentiating $\mathcal{D}$ with respect to time we can define a current,

$$J_x = \frac{1}{2i}\left(\dot{\psi}\nabla_y\psi^* - \text{c.c.}\right) \tag{2.87a}$$

$$J_y = \frac{-1}{2i}\left(\dot{\psi}\nabla_x\psi^* - \text{c.c.}\right) \tag{2.87b}$$

and thus we can say that $\mathcal{D}$ satisfies a continuity equation,

$$\partial_t\mathcal{D} + \nabla\cdot\boldsymbol{J} = 0. \tag{2.88}$$

Combining equations 2.81 and 2.88 we find that we can write $\mathcal{D}\boldsymbol{v} = \boldsymbol{J}$. In particular,

$$v_x = \frac{1}{2i\mathcal{D}}\left(\dot{\psi}\nabla_y\psi^* - \text{c.c.}\right) \tag{2.89a}$$

$$v_y = \frac{-1}{2i\mathcal{D}}\left(\dot{\psi}\nabla_x\psi^* - \text{c.c.}\right). \tag{2.89b}$$

A similar argument can be made for defect strings in three dimensions. The topological charge and the determinant become vectorial,

$$\boldsymbol{\rho}(\boldsymbol{x}) = \delta(\boldsymbol{\psi})\boldsymbol{\mathcal{D}}(\boldsymbol{x}) \tag{2.90}$$

$$\boldsymbol{\mathcal{D}} = \frac{1}{4i}\left(\boldsymbol{\nabla}\psi^* \times \boldsymbol{\nabla}\psi - \text{c.c.}\right), \tag{2.91}$$

thus obtaining the velocity as

$$\boldsymbol{v} = \frac{1}{\mathcal{D}^2}\boldsymbol{\mathcal{D}} \times \left(\dot{\psi}^*\boldsymbol{\nabla}\psi - \text{c.c.}\right), \tag{2.92}$$

where $\mathcal{D}^2 = \boldsymbol{\mathcal{D}} \cdot \boldsymbol{\mathcal{D}}$.

2.6.6.2 Numerical implementation

The numerical implementation of this method is relatively straightforward. However certain caution has to be taken when interpreting the data. We are reminded by equations 2.81 and 2.90 that our quantities of interest, namely $\mathcal{D}$ and $\boldsymbol{v}$ (and $\boldsymbol{\mathcal{D}}$ in the case of three dimensions), are *only defined at the location of the defect*. This point is of extreme importance, as failure to account for this fact when calculating these quantities renders the resulting velocity statistics meaningless.

In practice, to calculate the velocity statistics we proceed the following way:

- We calculate the determinant $\mathcal{D}$ (or $\boldsymbol{\mathcal{D}}$ in three dimensions). See Fig. 2.13 for a plot of the determinant in two dimensions, and Fig. 2.14 for a plot of the three-dimensional case.

- We calculate the components of $\boldsymbol{v}$, and if needed its magnitude v, normalized by $\mathcal{D}$ as prescribed in equations 2.89 or 2.92, making sure the use an adequate definition for the numerical first derivative (see eqs. 2.48 and 2.49), although the actual choice is not too critical. Also, we use $\dot{\psi}(t) \equiv \psi(t) - \psi(t-1)$, since in the Cell Dynamics System spirit the time differential is $\Delta t = 1$.

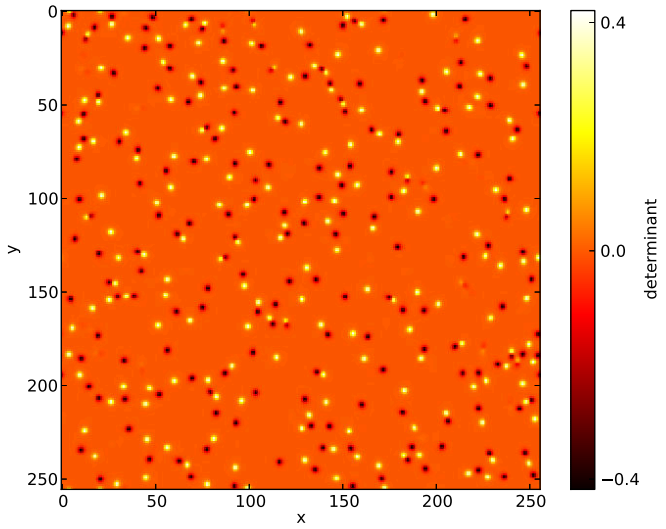


FIGURE 2.13: Two-dimensional representation of the determinant $\mathcal{D}$ after a critical quench. Color scale shows the range of values for $\mathcal{D}$. Most of the field has a value near zero, whereas the defects can be seen as highly localized maxima or minima of this field.

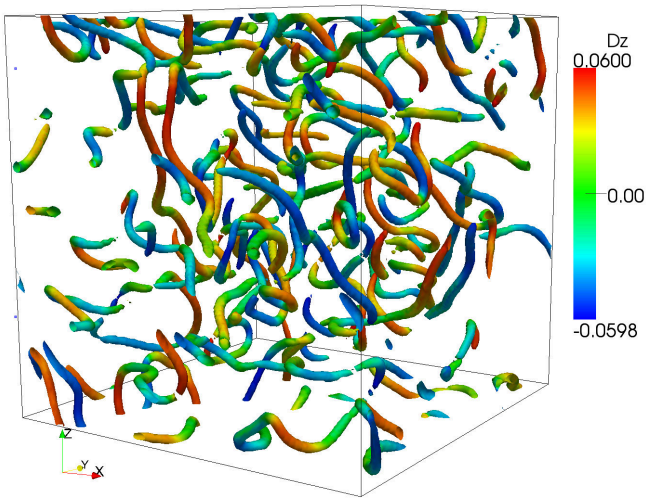


FIGURE 2.14: Three-dimensional representation of the determinant $\mathcal{D}$ and its z-component $\mathcal{D}_z$ after a critical quench. Isosurfaces enclose the points near the maxima of $|\mathcal{D}|$. Color scale shows the range of values for $\mathcal{D}_z$. Like in two dimensions, most of the field has a value near zero, whereas the strings can be seen as highly localized. We can notice that the maxima and minima of $\mathcal{D}_z$ tend to be aligned parallel or antiparallel with the z-axis, revealing the meaning of this component as the z-component of the vector of circulation about a line element of the string.

- Since $\mathcal{D}$ is also only defined at the location of the defects, we only retain the values of $\boldsymbol{v}$ at the locations where the magnitude of $\mathcal{D}$ is near its maximum. We normally used a value between 70 and 80% of the maximum of the magnitude of $\mathcal{D}$. For an example of the resulting data, see Fig. 2.15.

Strictly speaking, it is always necessary to normalize the distribution using the determinant, even though its value will be close to zero in most places. Again, this is not a problem, as we are only interested in

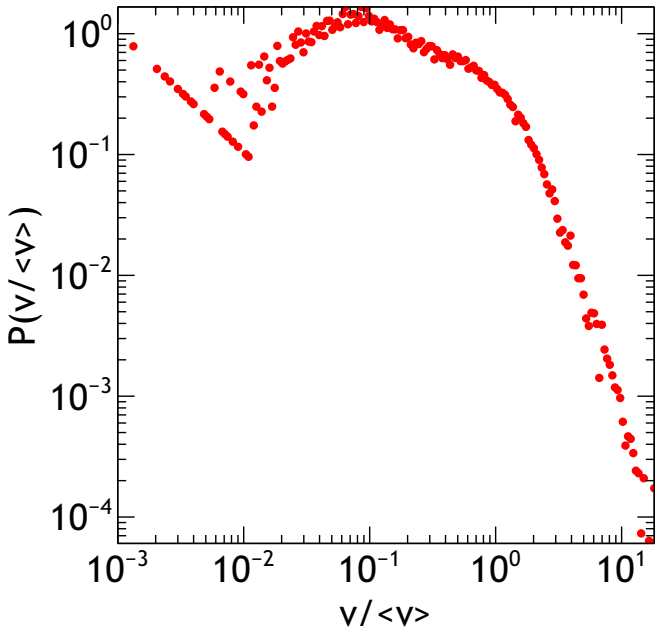


FIGURE 2.15: An example of the probability distribution of defect speeds for a two-dimensional system undergoing a quench, as calculated using a numerical implementation of eqs. 2.89. The data is an average obtained after running a simulation over 96 randomized initial conditions.

extracting values close to the defects, where the determinant's value deviates significantly from zero. Also, if we are normalizing the velocity distribution with its mean, we don't really need to normalize it using $\mathcal{D}$, although the efficiency gained by doing this is likely minor.

This method works with no problems as long as the topological defects we are interested in can be located by the zeros of the field, and are localized. For example, if we run the CDS superfluid simulation in a highly dissipative regime, or in a frozen defect phase in the Complex Ginzburg Landau sense [61, 62], where the defects behave as particles, then they can be localized with precision. But if we decrease the dissipation and the simulation is ran on a "defect turbulence" phase [61, 62], the defects cease to be as localized and they don't quite behave as particles and the zeros are shallower objects. See Fig. 2.16 for an example of this behavior.

It is also worth reminding that the definition of the hydrodynamic velocity using the Madelung transformation $\boldsymbol{v}_{\text{hydro}} = \boldsymbol{\nabla}\phi$ (where ϕ comes from the definition $\psi = Ae^{i\phi}$) is only valid away from the singularity. To help us in keeping us away from the defect, when taking statistics of $\boldsymbol{v}_{\text{hydro}}$ we also use $\mathcal{D}$ to extract the values of this velocity away from the defect, so as not to taint the high-speed values with artifacts from calculating too close to the defect core.

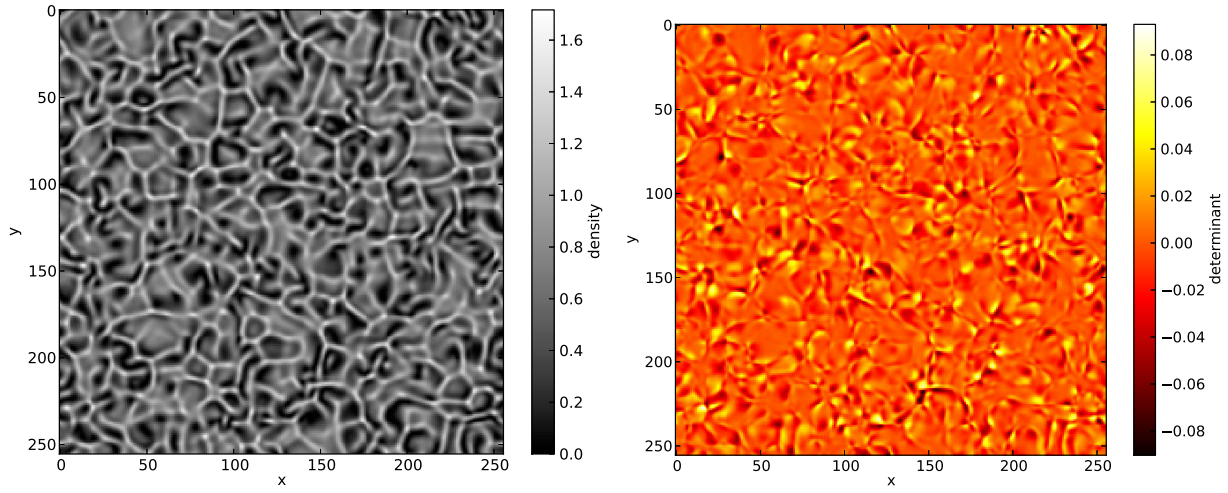


FIGURE 2.16: Direct integration of the complex Ginzburg-Landau equation in the "defect turbulence" regime, used to illustrate a case where the defect tracking method does not work. Left: Density field (the module squared of the complex field); the zeros of the field are not well localized, making it hard to pinpoint a "center." Right: Determinant $\mathcal{D}$ calculated on the field represented to the left; as expected, the lack of localization is reflected in the fact that the values of the determinant vary in a shallow way, making it useless to locate particle-like objects in this case.

2.6.6.3 Summary

In conclusion, this method allows us to extract velocity information for the defects directly from the definition of the order parameter. We can apply this method not only to our own model, but also to other more complex situations such as Swift-Hohenberg stripes (see Chap. 3 or Angheluta *et al.* [60]) or phase-field crystal defects, as long as the defects can be located using the zeros of the order parameter field. This also means that we will miss topological defects where the zeros are poorly localized, such as in the Complex Ginzburg-Landau wave turbulence regime, or defects of a different nature, not localizable by the zeros of the order parameter.

The fact that we can locate and localize the defects using the values of the determinant $\mathcal{D}$, means that we can use it in practice to numerically discriminate between the "inside" and the "outside" of the defect. This allows us to extract, for example, the values of the hydrodynamic velocity field using the Madelung transformation only where it is properly defined.

2.7 Conclusion

In this Chapter I introduced an efficient model for quantum fluids based on a Cell Dynamical Systems approach. It features both conservative and dissipative dynamics, and it allows the forcing of the system

by an external flow representing a normal fluid. The model was validated, as it exhibits topological defects and scaling behavior predicted for systems obeying the same physical attributes. Finally, I proposed an extension to the model that allows a fully coupled and efficient two-fluid model of superfluids.

This is yet another success of cell-based modeling, where the physics of the system are accounted for directly, instead of being represented as differential equations, and then these equations are simulated. This allows for more efficient numerical evolution and more convenient calculation of asymptotic quantities in the system.

Chapter 3

Anisotropic velocity statistics of topological defects under shear flow

We report numerical results on the velocity statistics of topological defects during the dynamics of phase ordering and non-relaxational evolution assisted by an external shear flow. We propose a numerically efficient tracking method for finding the position and velocity of defects, and apply it to vortices in a uniform field and dislocations in anisotropic stripe patterns. During relaxational dynamics, the distribution function of the velocity fluctuations is characterized by a dynamical scaling with a scaling function that has a robust algebraic tail with an inverse cube power law. This is characteristic to defects of codimension two, e.g. point defects in two dimensions and filaments in three dimensions, regardless of whether the motion is isotropic (as for vortices) or highly anisotropic (as for dislocations). However, the anisotropic dislocation motion leads to anisotropic statistical properties when the interaction between defects and their motion is influenced by the presence of an external shear flow transverse to the stripe orientation.

3.1 Introduction

The small-scale dynamics of interacting defects plays an important role in the evolution of complex systems. In particular, topological defects are a common occurrence in systems supporting a continuous symmetry that is spontaneously broken in the process of a non-equilibrium phase transition. One central question is how the universal properties and scaling laws near a critical phase transition relate to the presence and

interactions of defects. The formation and evolution of topological defects is typically formulated in the framework of the Ginzburg-Landau theory of symmetry-breaking phase transitions, where defects are described as phase singularities in a complex order parameter field (rotational symmetry) [1, 62].

The equilibrium structure of isolated topological defects is deeply rooted in their topological properties and is relatively well studied and understood [1]. In contrast, the dynamical and statistical properties of interacting topological defects during a non-equilibrium phase transition are far less understood, and are the subject of more recent systematic analyses. Numerous studies have focused on the statistical properties of topological defect ensembles in relation to the large-scale properties of the system. Examples range from the quenching dynamics during phase-ordering kinetics [12, 63], the motion of defects in convection patterns [64], the dislocation dynamics in crystal plasticity [65, 66] or the vortex filament motion in quantum flows [3]. A common characteristic of these apparently disparate systems is that they support codimension-two topological defects; that is dislocations and vortices which in a two-dimensional space (2D) become point defects, and defect filaments or loops in a three-dimensional space (3D). One common finding is the presence of a robust scaling law in the local velocity statistics for these kind of defects. Recent experiments on decaying quantum turbulence in ^{4}He report that the velocity field v induced by quantized vortices is characterized by a v^{-3} scaling, attributed to the rare reconnection events between vortex filaments [67] and reproduced numerically in atomic Bose-Einstein condensates [68] and counterflow turbulence [69]. Similar velocity statistics has been observed in a discrete dislocation dynamics model of crystal plasticity [70] and in experiments on thermal convection in an inclined fluid layer [71].

Theoretically, the asymptotic tail of the velocity probability distribution $P(v)$ can be calculated in a statistical formulation of random stationary configurations of point defects interacting through a logarithmic potential in a two dimensional space (2D) [65, 72]. The model predicts the same tail distribution both in neutral systems (zero net topological charge) and systems with a single-charge distribution. An inverse cubic scaling is consistent with the approximation of the nearest neighbor interaction between defects uniformly distributed in space [72]. In theoretical studies of defect motion during phase-ordering kinetics, the inverse cubic law is related to the annihilation events of defect loops or between point defects with opposite topological charges [12, 57, 73]. The coarsening during phase ordering is reflected in a time-dependent density of defects and their velocity distribution $F(v,t)$, which is characterized by a dynamical scaling law related to the growth law of the characteristic length scale in the ordering kinetics [57, 73] (see also [74]). For

non-conservative dynamics of the order parameter, the distribution of velocity for point defects in 2D takes the form $F(v,t) = \langle v(t)\rangle^{-1} P(v/\langle v(t)\rangle)$, where the scaling function is $P(x) \sim x(1+x^2)^{-2}$ and the ensemble average velocity $\langle v(t)\rangle$ at time t is related to the average distance between defects $L(t)$ at time t and scales with time as $\langle v(t)\rangle \sim 1/L(t) \sim t^{-1/2}$ [57, 73]. A different scaling exponent for the scaling function $P(x)$ is predicted for defect filaments in three-dimensions (3D) [12, 73], whereas experiments [67] and numerics [69, 70] suggest the same scaling as for point defects.

In contrast to isotropic vortex dynamics, dislocation motion in crystals, as well as in stripe patterns, is typically anisotropic when confined to certain gliding and climbing planes. In addition, dislocations often coexist and interact with other kinds of defects such as disinclinations and grain boundaries, which makes it harder to study in isolation. For this reason, phase ordering is much more difficult to study in isotropic stripe phases and polycrystalline phases, then in anisotropic stripes and single crystals where only dislocations are present [75, 76].

Stripe ordering is a common pattern occurring in a diversity of systems from the zebra patterns to sand ripples and in classical fluid convection systems, where defects are local tears of the underlying pattern [64]. Anisotropic stripes or rolls develop by an uniaxial ordering of stripes as happens, for instance, in electrically driven convection flows of nematic liquid crystals (electrohydrodynamic convection) [77], or in thermal convection flows of isotropic fluids down an inclined plane [78].

During relaxational dynamics, where the motion of defects is dominated by mutual interactions prior to annihilations, the statistics of the velocity components keeps the same form even in the presence of strongly anisotropic motion of defects. A numerical study of 2D phase ordering after a quench from a disordered state described by a non-conservative time dependent real Ginzburg-Landau model showed that the isotropic motion of point vortices is characterized by a statistical distribution with an inverse cubic tail in the scaling function $P(v)$, as predicted theoretically [55]. Similar statistical distributions for the climbing (motion along the direction of the stripes) and gliding (motion across the stripes) velocities of point dislocations have been reproduced in a numerical study of phase ordering in anisotropic stripes in two dimensions [59]. This is consistent with the theoretical understanding that the dislocation motion in anisotropic stripes can be in fact mapped onto a Ginzburg-Landau vortex-dynamics [1]. This also means that, to the leading order approximation, the interactions between dislocations are expected to be similar to those between vortices.

The statistics of defect motion during non-relaxational evolution of the system, self-sustained or driven

by an external field, is less well-understood due to correlation effects or additional driving forces apart from the mutual nearest neighbor interactions between defects. A self-sustained motion of defects is obtained in convection patterns when the mean flow due to vertical vorticity, driven by the undulations in the normal stripes and the presence of defects, acts as a self-induced drift in the motion of defects [79]. This non-trivial dynamics of defects leads to a spatio-temporal chaotic state also know as "defect turbulence," which was observed experimentally in fluid convection systems [78, 80] or diffusion-reaction systems [81], and studied in numerous theoretical and numerical investigations [56, 79, 82–84]. In this chaotic dynamical regime, a statistically stationary distribution of the number of defects is maintained by the defect annihilations and the spontaneous creations of pairs due to the phase instability. To leading order in the approximation of well mixed and independent defects, the distribution of the number of defects follows the Poisson statistics with mean square fluctuations given by the mean number of defects [82]. The well mixed assumption implies that defect pairs are being created and annihilated randomly, whereas experiments and numerical studies suggest that more often defects created in a pair at a given time tend to annihilate with each other in the same pair at a subsequent time [56, 81], which means that correlations between defects are important effects in their creation/annihilation dynamics and leads to a modified Poisson statistics in their number fluctuations [84]. A theoretical understanding for the effect of the self-induced mean flow on the collective statistical properties of defect motion is still lacking. However, experimentally measured velocity statistics during the spatio-temporal chaotic dynamics of uniaxial stripes in inclined layer convection are observed to be slightly anisotropic and the exponents in the tail distribution of both climb and glide motion are close to -3 [71]. This is suggestive of a dynamical regime dominated by annihilations of dislocation pairs.

In this chapter, we consider a simpler setup where non-relaxational motion is driven by an externally imposed flow such that defects are constantly created and annihilated leading to a statistically stationary defect dynamics. This can be attained in an anisotropic stripe system when a shear flow is acting normal to the stripe orientation. The role of the shear flow is different from the commonly studied case of shear alignment of isotropic stripes [71, 85, 86] or the buckling instability under shear acting along unixial stripes orientation [87].

The purpose of this chapter is threefold: i) to present an efficient numerical method for tracking the position and velocity of topological defects and ii) its application to study the collective motion of dissipative vortices both in 2D and 3D and dislocations in 2D; iii) to report on numerical results where the anisotropic

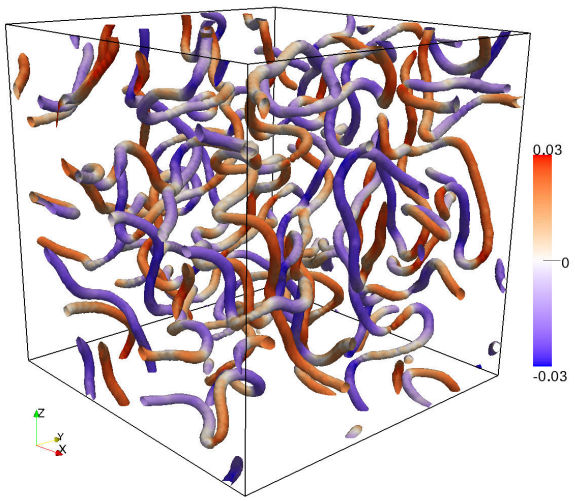


FIGURE 3.1: Snapshot of a measure of the charge density vector field for a configuration of vortex filaments in 3D simulation of phase ordering. The figure shows the x-component $\mathcal{D}_x$ of the Jacobian determinant defined in Sec. 3.2. The system size in this simulation is 128^3.

motion (glide and climb) of dislocations subjected to a simple shear flow is explicitly manifested in the velocity distributions, even though the statistics during phase ordering are similar to those corresponding to the isotropic motion of vortices.

To track the position and velocity of topological defects, we implemented a numerical method inspired by analytical treatments of Halperin [11] and Mazenko [57]. The method was originally developed to locate defects in an $O(2)$-symmetric order parameter with a Ginzburg-Landau relaxation dynamics in 2D. We show numerically that this method works very well for Ginzburg-Landau dynamics both in 2D and 3D and it is also suitable for tracking dislocations in systems controlled by anisotropic Swift-Hohenberg dynamics. Measuring the velocity statistics of vortices during relaxational dynamics, we find a universal inverse cubic tail for defects of the same codimension, that is point vortices in 2D and vortex filaments in 3D. The scaling law is directly related to the pairwise interactions between vortices prior to annihilation and reconnection events (in 3D). Finite size core effects induce a Gaussian cut-off to the v^{-3} scaling. A similar statistical behavior is observed in the velocity of dislocations in anisotropic stripe patterns. Despite the fact that dislocations are dominated by their transverse motion, and thus are highly anisotropic, the distribution of the climb and glide velocities shows the same long tail behavior. In the presence of an external shear flow that leads to non-relaxational dynamics, the motion anisotropy is explicitly manifested in different statistics of the velocity components. While the slow motion is highly influenced by the shear flow, the high speed

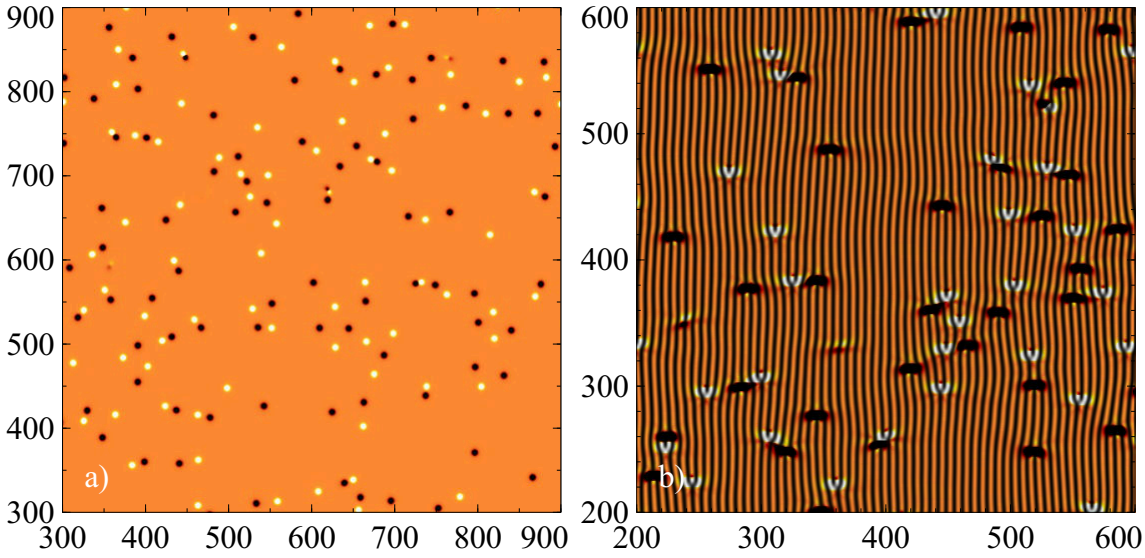


FIGURE 3.2: (a) Snapshot of the charge density field corresponding to a configuration of point vortices in 2D simulations. In panel (b), we show the dislocations in an underlying anisotropic stripe configuration. The lighter blobs correspond to defects that have a positive charge, while the darker blobs are the defects of the opposite charge. The system size in both cases is 1024^2, while the snapshots are drawn from a subset.

limit may still be dominated by the nearest neighbor interactions.

The chapter is organized as follows. Following this introduction, we proceed in Sec. 3.2 to discuss a method of efficiently tracking topological defects and apply it to a collection of vortices as well as ensembles of dislocations. Sec. 3.3 presents numerical results on the vortex velocity statistics in 2D and 3D simulations of phase ordering. We discuss the statistics of dislocations during phase ordering and non-relaxation dynamics sustained by an external shear flow in Sec. 3.4. Concluding remarks and summary are provided in Sec. 3.5.

3.2 Defect dynamics

Here we present a numerically efficient method for tracking codimension-2 topological defects. The method is applied to Ginzburg-Landau dynamics of vortices in 2D and 3D, as well as to Swift-Hohenberg dynamics of dislocations in 2D anisotropic stripes. The effect of hydrodynamic interactions in the presence of an external shear flow is discussed in the context of dislocation dynamics.

3.2.1 Locating and tracking of defects

The identification and evolution of a large population of defects is generally a non-trivial problem in systems described by continuum approaches. Moving from the field variables to the discrete particle variables is not straightforward. In most defect studies, one resorts to various approximate methods to estimate the locations and velocity of defects by following their trajectories [55, 71, 76].

In systems that can be described by an $O(2)$-symmetric order parameter $\psi(\boldsymbol{r}, t)$ whose evolution depicts the ordering kinetics from an initially disordered state to an ordered state (either isotropic and homogeneous or a periodic pattern), one can use an elegant method based on a transformation from the order parameter dynamics to the discrete defect dynamics. The analytical formulation of this method was pointed out first by Halperin [11], and was subsequently extended by Mazenko to determine the velocity of various topological defects [12, 55, 57, 58]. To our knowledge, this method has not been previously implemented numerically. We show that it is an efficient numerical tool used for tracking the evolution of various types of defects.

The basic idea of this technique is that topological defects are located at the zeroes of the complex order parameter field $\psi(\boldsymbol{r}, t)$ [11]. The transformation from field to particle variables is determined by the Jacobian determinant $\mathcal{D}(\boldsymbol{r}) = \|\partial\psi_n/\partial r_j\|$, where $n = 1, 2$ stands respectively for the real and imaginary components of the order parameter field, i.e. $\psi = \psi_1 + i\psi_2$, and $j = 1, \ldots d$, with d being the spatial dimension. Thus, for $d = 2$, $\mathcal{D}(\boldsymbol{r})$ is a scalar quantity. Its sign determines the topological charge, i.e. $q = \mathcal{D}(\boldsymbol{r})/|\mathcal{D}(\boldsymbol{r})| = \pm 1$, and the charge density is given as

$$\rho(\boldsymbol{r}, t) = \delta(\psi)\mathcal{D}(\boldsymbol{r}, t) = \sum_{i=1}^{N} q_i \delta(\boldsymbol{r} - \boldsymbol{r}_i), \tag{3.1}$$

for a collection of N point vortices. The extension to string defects in $d = 3$ is that the Jacobian determinant becomes a vector field $\mathcal{D}_j$(r) related to the vortex filament density by

$$\rho_j(\boldsymbol{r}, t) = \delta(\psi(\boldsymbol{r}, t))\mathcal{D}_j(\boldsymbol{r}), \tag{3.2}$$

where the notation for the Dirac δ distribution is used [12].

The defect velocity $\boldsymbol{v}$ is determined from the property of topological defects that their total charge is

conserved (defects are created and annihilated in pairs of opposite charge), namely

$$\partial_t \rho + \nabla \cdot (\rho \boldsymbol{v}) = 0, \tag{3.3}$$

with the charge density $\rho(\boldsymbol{r}, t)$ defined above. For example, in the case of point defects in 2D, the Jacobian determinant becomes $\mathcal{D} = 1/(2i)(\nabla_x \psi^* \nabla_y \psi - \nabla_x \psi \nabla_y \psi^*)$, where ψ^* is the complex conjugate of ψ. By differentiating $\mathcal{D}$ with time, a current $\boldsymbol{J}^{(\dot{\psi})}$ can be defined as [57]

$$\boldsymbol{J}_\alpha^{(\dot{\psi})} = -\frac{i\epsilon_{\alpha\beta}}{2}(\dot{\psi}\nabla_\beta \psi^* - \dot{\psi}^* \nabla_\beta \psi), \tag{3.4}$$

such that the $\mathcal{D}$-field satisfies the continuity equation

$$\partial_t \mathcal{D} + \nabla \cdot \boldsymbol{J}^{(\dot{\psi})} = 0. \tag{3.5}$$

Summation over repeated indices is implied and $\epsilon_{\alpha\beta}$ is the two dimensional antisymmetric tensor, $\epsilon_{xx} = \epsilon_{yy} = 0$ and $\epsilon_{xy} = -\epsilon_{yx} = 1$. From Eqs. 3.3 and 3.5, the defect velocity is determined as $\boldsymbol{v} = \boldsymbol{J}^{(\dot{\psi})}/\mathcal{D}$. The defect velocity depends on the dynamics of the order parameter through its time derivative $\dot{\psi}(\boldsymbol{r}, t)$. Explicitly, the velocity components are given by

$$\begin{aligned} v_x &= -i\frac{\dot{\psi}\nabla_y \psi^* - \dot{\psi}^* \nabla_y \psi}{2\mathcal{D}}, \\ v_y &= i\frac{\dot{\psi}\nabla_x \psi^* - \dot{\psi}^* \nabla_x \psi}{2\mathcal{D}} \end{aligned} \tag{3.6}$$

where ψ^* is the complex conjugate of ψ-field and $\dot{\psi}$ is the time derivative of ψ which determines the evolution of the order parameter. This can be generalized to $d = 3$, in which case the velocity of vortex filaments is calculated as [12]

$$\boldsymbol{v} = \frac{\boldsymbol{\mathcal{D}} \times (\dot{\psi}^* \nabla \psi - \dot{\psi} \nabla \psi^*)}{2\mathcal{D}^2}, \tag{3.7}$$

where $\mathcal{D}^2 = \sum_{j=1}^{3} \mathcal{D}_j \mathcal{D}_j$ and the velocity vector field is $\boldsymbol{v} = (v_x, v_y, v_z)$.

In the dilute defect density limit, it can be shown that vortex velocity defined by Eq. 3.6 becomes a function of the phase and amplitude gradients of the order parameter ψ near the vortex core [58]. The formula is exact and applies equally well for a high density of defects.

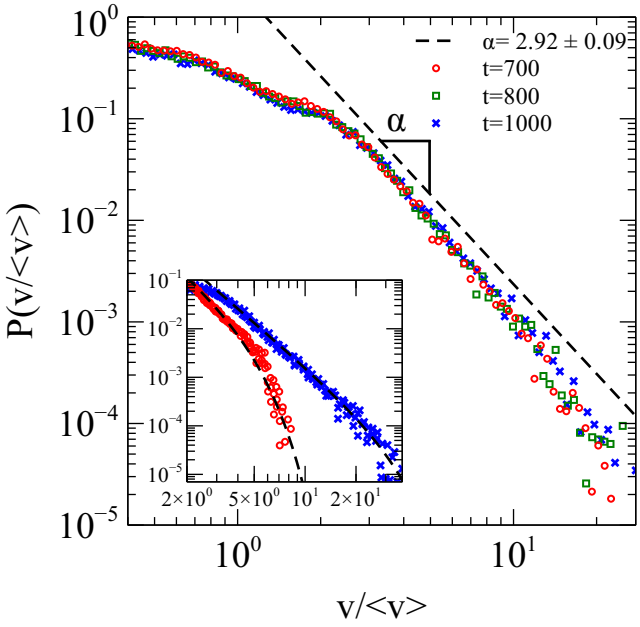


FIGURE 3.3: Collapsed probability distribution function (scaling function) of absolute velocity of point vortices in 2D during phase ordering. (Inset) PDF of the defect velocity for different core sizes (open circles correspond to larger core size and crosses correspond to smaller core size) to show that the Gaussian cut-off depends effectively on the vortex core size. The model parameters for the inset figure are $A = 2.05$ (open circles), $A = 1.05$ (crosses) and $C = 3/20(1 + A)$. In the main graph, $A = 1.5$. Here $v \equiv |\mathbf{v}|$.

3.2.2 Application to vortices

Vortices are defined as the zeros of an order parameter $\psi(\boldsymbol{x}, t)$ with rotational symmetry (complex field) [1]. The fact that the complex field vanishes at the core of a defect is equivalent to a phase singularity, i.e. the phase of the order parameter varies discontinuously around a closed contour surrounding the defect. The phase θ is obtained from $\psi = |\psi|e^{i\theta}$. The shift in phase around the contour or the winding number, i.e. $\oint \nabla\theta \cdot d\boldsymbol{l} = 2\pi n$, defines the topological charge of the defect. A single vortex corresponds to a unit of topological charge, that is $n = 1$.

We now consider the nonconservative evolution of a $\psi(\boldsymbol{r}, t)$-field described by the time dependent Ginzburg-Landau equation given by

$$\partial_t \psi = \nabla^2 \psi + \psi(1 - |\psi|^2), \tag{3.8}$$

which we simulate both in 2D and 3D. For computational efficiency, we solve Eq. 3.8 by a cell dynamical system (CDS) algorithm, that was originally developed for studying spinodal decomposition dynamics [33] and extensively used to study phase ordering of systems with continuous symmetry [5, 6, 63]. In the Ap-

pendix we provide a detailed description and recapitulation of the algorithm, for completeness, and define the parameters of the simulation used below. In particular, the depth of the quench corresponds to the parameter A, and the strength of the diffusive couplings in the model are denoted by C. Simulations in 2D are done on a system size of 1024^2 cells, while in 3D we use 128^3 cells for $dx = 1$. Unless otherwise noted, the values for the CDS parameters A, the depth of the quench, and C, the strength of the spatial coupling, are $A = 1.5$, $C = 3/20(1 + A)$ (for 2D), and $C = 3/24(1 + A)$ (for 3D). Results were averaged over 48 random initial conditions, unless otherwise noted.

The vortex dynamics from the Ginzburg-Landau evolution in Eq. 3.8 is similar to a previous one reported in Qian & Mazenko [55]. Here, we use a different tracking method for locating the defects and extend the analysis to vortex filaments in 3D.

A snapshot of the charge density field for vortex filaments in 3D obtained using Halperin and Mazenko's method, discussed in the previous section, is shown in Fig. (3.1). The charge density field is directly proportional to the $\mathcal{D}$-field, which is zero everywhere except along the vortex filaments. A similar representation is obtained for point vortices in 2D, where the charge field is localized at the vortex core and vanishes everywhere else as shown in Fig. (3.2 (a)). Since the charge density is directly related to the $\mathcal{D}$-field, it means that the defect velocities are meaningfully defined only at the defect positions. In the numerical discretizations, defects are associated with small blobs (in 2D) or thin tubes (in 3D) with a specific characteristic size that defines the vortex core size. We define the defect regions as the locations at which the absolute charge density is above 75% of the theoretical value of $|q| = 1$. The values of the $\mathcal{D}$-field are finite within these regions and thus the division is also finite. The velocity of the located defects is determined from Eq. 3.7 for filaments in 3D, and Eq. 3.7 for point vortices in 2D. The time derivative $\dot{\psi}$ of the ψ-field is defined by the right hand side expression in Eq. 3.8, namely $\dot{\psi} \equiv \nabla^2\psi + \psi(1 - |\psi|^2)$.

3.2.3 Application to dislocations

Here, we focus on tracking dislocations in anisotropic stripe patterns. A similar tracking method can be extended to locate the defects in a crystal phase, and will be the subject of a separate study reported elsewhere.

We consider the defects in a periodic pattern characterized by a preferred wavenumber $\boldsymbol{k}$ formed by stripes. Stripe patterns occur in a variety of systems, typical examples being the convective rolls in Rayleigh Bénard convection of isotropic fluids [64] or convective flows in the nematic liquid crystals [88], in the dy-

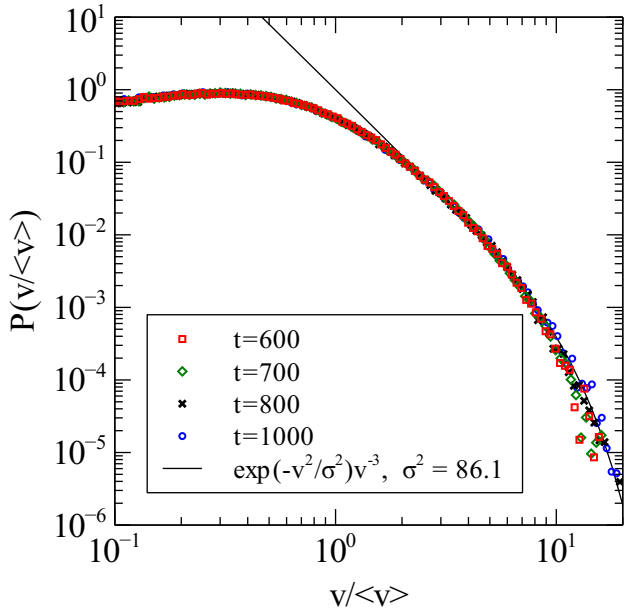


FIGURE 3.4: The scaling function of the probability distribution function of the absolute velocity of vortex filaments in 3D during phase ordering. Here $v \equiv |\mathbf{v}|$.

namics of diblock copolymers [89], etc. When the orientation of the local ordering is random, as in isotropic stripes, we encounter both isolated defects such as dislocations or disinclinations as well as grain boundaries. The coexistance of different types of defects makes it difficult to analyze their statistics. Moreover, the ordering kinetics of isotropic stripes tends to be dominated at large times by grain boundary slow motion, which can lead to glassy configurations [75]. By fixing the orientation of the stripes along a preferred axis, point defects such as dislocations can be isolated from the other types of defects. Examples of anisotropic stripes are in electrohydrodynamic convection of planary aligned liquid crystals [80], or Rayleigh-Bénard convection of an inclined fluid layer [71].

The statistics of dislocations in anisotropic stripes has been discussed previously by Qian & Mazenko [59], where they propose a model based on an effective Swift-Hohenberg (SH) free energy with an additional term that accounts for the coupling to an external field aligning the stripes along a preferred direction. The stripe pattern in 2D is represented by a real periodic field $u(\boldsymbol{r}, t)$, which satisfies an anisotropic SH-dynamics given by

$$\partial_t u = (1 - r|u|^2)u - (1 + \nabla^2)^2 u - c\nabla_x^2 u, \tag{3.9}$$

where the last term is added to impose a preferred orientation of the stripes along the vertical y-axis with

$c > 0$ being the coupling strength to the external field. The quench depth $r > 0$ is interpreted in the context of convection patterns as the deviation from the onset of convection, $r \approx R/R_c - 1$, where R is the Rayleigh number and R_c is the critical R at the onset [64]. The anisotropic preferred orientation can be seen by linearizing Eq. 3.9 around the mode solution $u \sim \exp(\omega t + ik_x x + ik_y y)$ with the growth rate obtained from Eq. 3.9 as

$$\omega = (1 - r) - (1 - k_x^2 - k_y^2)^2 + ck_x^2, \tag{3.10}$$

and imposing the condition $\omega(r; k_x, k_y) = 0$ at the onset of instability with respect to a mode of wavenumbers k_x and k_y. The condition is found by minimizing ω with respect to k_x and k_y, i.e. $\partial\omega/\partial k_x = 0$ and $\partial\omega/\partial k_y = 0$. This leads to $k_y = 0$ and $k_x = \sqrt{1 + c/2}$ for $c > 0$.

We consider the amplitude formulation of Eq. 3.9 with an additional contribution due to an external shear flow. We impose a shear flow that is normal to the main orientation of the stripes to allow for the nucleation of defects due to wavenumber shifts by shear deformation. In order to track dislocations in a complex order parameter field, we write the periodic field in terms of its complex envelope field $\psi(\boldsymbol{r}, t)$, namely $u(\boldsymbol{r}) = \sqrt{r}\psi(\boldsymbol{r})e^{i\boldsymbol{k}\cdot\boldsymbol{r}} + c.c.$. Without loss of generality, we consider stripes with the wavevector parallel to the horizontal x-axis, i.e. $\boldsymbol{k} = (k_0, 0)$. The complex ψ-field satisfies an amplitude equation derived from Eq. 3.9 and given to the leading order in r as

$$\partial_t\psi + \dot{\gamma}_x\mathcal{L}_x[\psi] = r(1 - |\psi|^2)\psi - \mathcal{L}^2[\psi] - c\mathcal{L}_x^2[\psi], \tag{3.11}$$

where $\mathcal{L} \equiv (\nabla^2 + 2i\boldsymbol{k}\cdot\nabla)$ is derived from $(1 + \nabla^2)$ and $\mathcal{L}_x \equiv \nabla_x + ik_0$ comes from the gradient ∇_x. The advection term is determined by a velocity field, which hereby is taken as a simple shear flow $\dot{\boldsymbol{\gamma}} = v_0 y\hat{\boldsymbol{x}}$, and the last term is added to impose a preferred orientation of the stripes long the vertical y-axis.

We integrate numerically Eq. 3.11 using a 4^{th}-order Runge-Kutta scheme and a spherical approximation for the gradients [38] (see Appendix) on a square domain of size $1024dx \times 1024dx$. The time step is $dt = 0.05$ and the spatial resolution is $dx = \pi/4$ so that about 8 grid points are used to resolve the pattern wavelength $\lambda = 2\pi/k_0$, with $k_0 = 1$. The other parameters are set to $c = 1$, $r = 1$ and v_0 is a changing parameter. In the absence of shear, period boundary conditions on all sides are used. At a finite shear rate, we impose a zero flux boundary conditions of the upper and lower boundaries and periodic conditions on the lateral boundaries.

Dislocations are efficiently located as the zeros of the complex envelope field ψ using Mazenko's algorithm. In Fig. (3.2 panel (b)), we illustrate a stripe configuration with the location of dislocations and their topological charge proportional to the Jacobian determinant $\|\partial\psi_n/\partial r_j\|$. The velocity of dislocations is obtained using Eq. 3.6 with the evolution of the order parameter given by the right hand side of Eq. 3.11, namely $\dot{\psi} \equiv r(1-|\psi|^2)\psi - \mathcal{L}^2[\psi] - c\mathcal{L}_x^2[\psi]$.

3.3 Vortex Statistics

To determine the velocity statistics of vortices, we initiate the system in a disordered state and follow the ordering kinetics dominated by the initial formation and subsequent coarsening of topological defects. At a particular time, we calculate the local defect velocities $\boldsymbol{v}$ using Mazenko's method as described above. We save the absolute values, $v = |\boldsymbol{v}|$, every few time iterations and run the system from 48 random initial conditions. This way, we compute the probability distribution function of the defect velocity at a given time, i.e. $F(v,t)$. In the asymptotic limit $t \to \infty$ of the coarsening dynamics, we expect scale invariance of the typical coarsening length scale, i.e. $L(t) \sim t^{1/2}$ for non-conservative dynamics (apart from logarithmic corrections in 2D). Hence the typical velocity obtained as $1/L(t)$ scales with time as $\langle v(t)\rangle \sim t^{-1/2}$ and corresponds to the velocity of defects in a pair prior to annihilation and separated by a distance of the order of $L(t)$. In the simulations, $\langle v(t)\rangle$ is the ensemble average velocity at a particular time, and when calculated over long times it converges to the expected asymptotic scaling. This dynamical scaling of the mean velocity implies also a scaling with time of $F(v,t)$. We notice that the time dependence in the PDF's can be eliminated by rescaling the velocity variables by their ensemble average values at a given time, i.e. $\tilde{v} \equiv v/\langle v(t)\rangle$. Analytically, this corresponds to the rescaling $F(v,t) = t^{1/2}P(vt^{1/2})$.

The scaling function $P(x)$ of the velocity distribution is a function of the rescaled velocity field $v/\langle v(t)\rangle$ and has a broad tail with an inverse cubic decay. This is shown in Fig. (3.3) for 2D simulations and Fig. (3.4) for 3D dynamics. The v^{-3} tail corresponds to the regime of large velocities obtained in pair interactions prior to annihilation events or, for 3D, also reconnections events. The -3 scaling exponent is determined by the logarithmic mutual interaction potential as shown in e.g. Refs. [73, 90]. Since point defects in 2D and filaments in 3D are both codimension-2 topological defects with the same type of interactions, we expect a similar scaling behavior. This is consistent with other numerical studies that also show that the tail of $P(v)$ distribution is dominated by the v^{-3} scaling both in 2D and 3D simulations [69, 70]. We provide additional

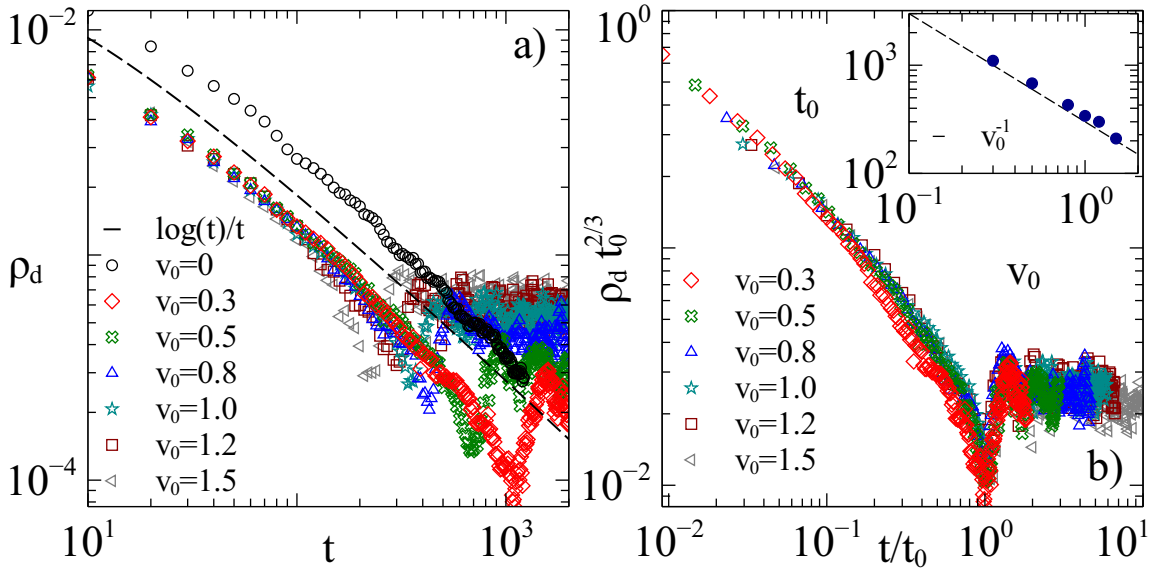


FIGURE 3.5: a) Density of defects ρ_d as a function of time t for different values of the imposed shear velocity v_0. b) Data collapse of the rescaled density with the mean density in the statistical steady state $\rho_0 \sim t_0^{-2/3}$, where t_0 is the cross-over time to the steady state. Inset figure shows how t_0 scales with v_0 as $t_0 \sim 1/v_0$.

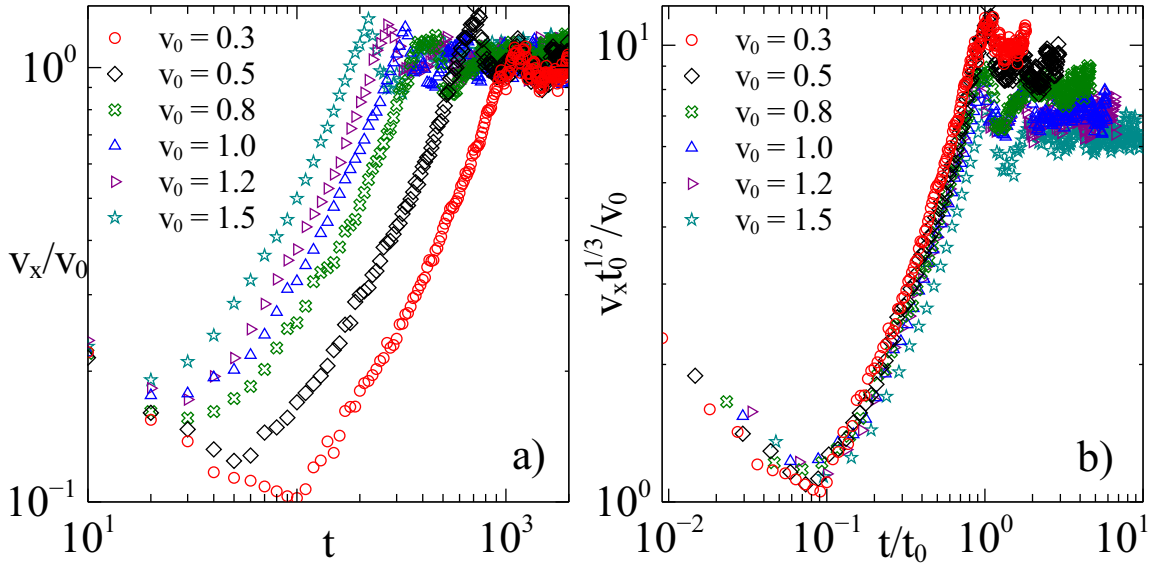


FIGURE 3.6: a) Evolution with time of the ensemble average of the velocity component $v_x \equiv \langle |v_x(t)| \rangle$ for different values of the shear rate. b) Data collapse of the rescaled v_x as a function of the rescaled time t/t_0.

numerical evidence for this scaling regime also during phase ordering kinetics.

From Figs. (3.3) and (3.4), we notice that the vortex core structure has a drastic influence on the defect velocity statistics [90]. This effect is given by a Gaussian tail which takes over the v^{-3} regimes at very large fluctuations. The vortex core size is more pronounced in 3D simulations, due to numerical bounds on higher spatial resolutions. It is easier to vary the aspect ratio between system size and vortex core size in 2D

simulations to observe the vortex core effect. In CDS simulations, the core size is fixed by the parameter A. We consider two different values of A, namely $A = 1.05$ for small cores and $A = 2.05$ for larger cores, and averaged over 2000 random initial conditions. The system size was reduced to 128^2 cells. The dependence of the Gaussian-cutoff on the core size is shown in the inset of Fig. (3.3), while the velocity distribution for an intermediate core size, parametrized by $A = 1.5$, is plotted in the main graph of Fig. (3.3), using 1024^2 cells and averaging over 48 initial conditions.

3.4 Dislocation statistics

Next we discuss the statistics of dislocations in anisotropic stripes during phase ordering and non-relaxational dynamics assisted by a shear flow.

3.4.1 Dynamical scaling regimes

In the early stages, the evolution of defects is dominated by pairwise interactions leading to annihilations and a decrease in the density of defects. We measure the density of defects $\rho_d(t)$ as the ratio between the effective area occupied by the defects and the total system area. Alternatively, the number of defects N can be estimated as the area occupied by the defects divided by the approximate core area of a defect. In simulations, we choose a large system size of 1024^2 and calculate $\rho_d(t)$ as a function of time, starting from a random initial condition. During the relaxation dynamics, the defect density for a very large system is equivalent to the averaged density over many initial conditions for a smaller system size, and is a smooth function of time [91]. As expected in the coarsening regime, the density obeys a power law in time as $\rho_d(t) \sim \log(t)/t$ like the density of vortices in Ginzburg-Landau theory [55]. This behavior is shown in Fig. (3.5a) by the data in open circles and is consistent with formal arguments that the anisotropic Swift-Hohenberg dynamics can be mapped onto the isotropic Ginzburg-Landau dynamics [1]. In Fig. (3.5a), we also plot the density $\rho_d(t)$ as a function of time for various values of the shear rate v_0. We notice that in the late stages, when long-range hydrodynamic interactions set in, the density of defects ceases to decrease monotonically, and approaches instead a statistically steady state. In this steady state , the defect density fluctuates in time about a mean value because of the sporadic pair creations and subsequent annihilations of defect pairs. Averages over initial conditions would lead to a constant mean density ρ_0 in this steady state, i.e. $\langle \rho_d(t) \rangle_{IC} \to \rho_0$ as $t \to \infty$. We also observe that the mean number of defects $\langle N \rangle$ in the steady state

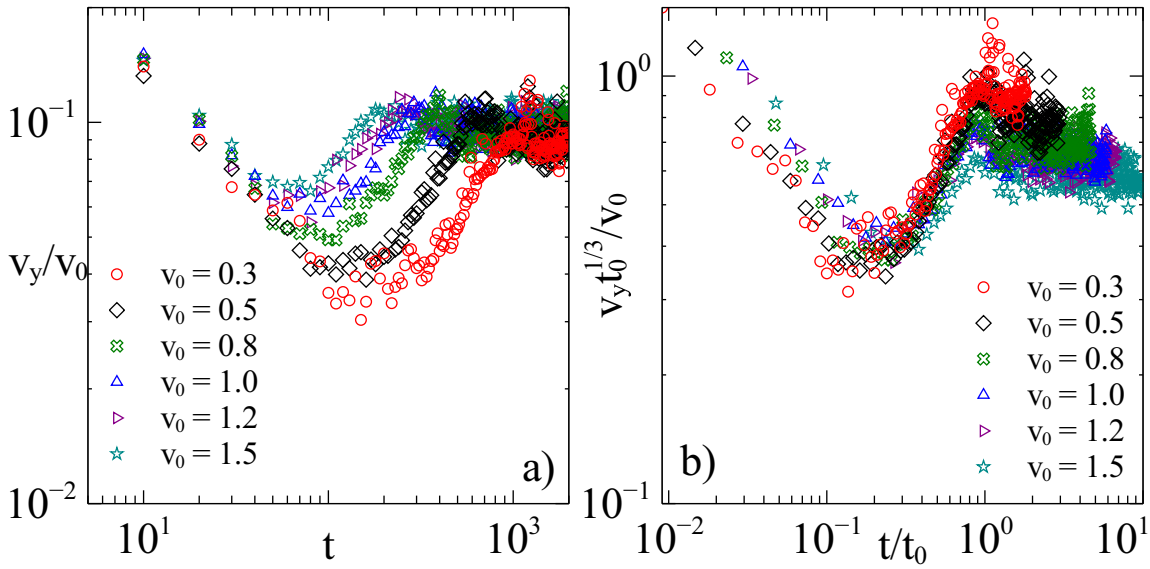


FIGURE 3.7: a) Evolution with time of the ensemble average of the velocity component $v_y \equiv \langle |v_y(t)| \rangle$ for different values of the shear rate. b) Rescaled v_y plotted as a function of the rescaled time t/t_0.

appears to increase with the applied shear v_0, suggesting that the creation rate of defect pairs depends on v_0. Also, the mean square fluctuations $\langle N^2 \rangle - \langle N \rangle^2$ in the number of defects increases monotonically with $\langle N \rangle$ with significant deviations from what would be expected in a Poisson process. It appears that the number statistics of defects behaves in a similar manner to that in defect turbulence [56], although a detailed analysis of this suggestion would be beyond the scope of this chapter.

From the data presented in Fig. (3.5b), we can determine the cross-over time t_0 to the statistical steady state, and we find that it increases to a first approximation as $t_0 \sim v_0^{-1}$, as shown in the inset of Fig. (3.5b). It turns out that phase ordering assisted by hydrodynamic effects slows down the growth of the typical distance between defects with corrections that follow a $L(t) \sim t^{1/3}$ law until saturating to the steady state. This implies that the steady state defect density can be estimated as $\rho_0 \sim L(t_0)^{-2} \sim t_0^{-2/3}$. Equivalently, the mean density in the steady state increases with the applied shear as $\rho_0 \sim v_0^{2/3}$. As a first step to see whether there is any data collapse associated with this scaling behavior, we plot the rescaled ρ/ρ_0 versus the rescaled t/t_0 as shown in Fig. (3.5b).

Because the normal stripes are aligned along the y-direction, the climb and glide motions of the dislocations correspond respectively to the vertical and horizontal velocities. The motion is anisotropic and dominated by the transverse (gliding) dynamics of dislocations. We observe that climb motion typically occurs when two dislocations of opposite topological charge approach each other to annihilate, otherwise

dislocations would move by gliding across the stripes. Nevertheless, when the two motions are driven by the local pairwise interactions between dislocations, they exhibit similar characteristics. In the early states, where the dynamics is controlled mainly by the phase ordering and annihilations of dislocations, the ensemble average of the net velocities (absolute values), $\langle|v_x|\rangle(t)$ and $\langle|v_y|\rangle(t)$, scale with time as $t^{-1/2}$ until hydrodynamic effects set in and the defects are accelerating until they cross over at t_0 to a statistically stationary state. This behavior is shown in Fig. (3.6a) for $\langle|v_x|\rangle$ and in Fig. (3.7a) for $\langle|v_y|\rangle$ corresponding to different values of the imposed shear velocity v_0. In the later stages, hydrodynamic effects become important and there is a transient regime where the mean velocities increase almost linearly with time until t_0 after which a statistically stationary state is reached. Since, we run large scale simulations for a given realization without averaging over initial conditions, the ensemble-averaged velocities, $\langle|v_x|\rangle(t)$ and $\langle|v_y|\rangle(t)$, correspond to time series in the statistical steady state. Averaging these fluctuations over many initial conditions would smooth out the late-stage time dependence to a constant value.

The mean square fluctuations in the steady state are an increasing function of the applied shear. We observe that the mean value of climb velocity is approximately an order of magnitude smaller than the typical velocity of gliding. A rescaling of velocity components as a function of the rescaled time in the units of t_0 is presented in Figs. (3.6b) and (3.7b), which however gives a poor data collapse. Improving it turned out to be a challenging task, one of the reasons being that there is an additional characteristic timescale given by the cross-over from the relaxation dynamics to the transient period of acceleration prior to the steady state. It may be that this timescale also plays a role in the scaling function, but we have not succeeded in including it to our satisfaction. This is an unresolved issue that deserves a separate detailed study.

The statistically stationary regime is characterized by fluctuations in the density of defects and their velocities around a mean value that depends on the imposed shear flow. Fluctuations in the defect density are attributed to sudden nucleation of dislocation pairs and their subsequent annihilation, either with the same pair member, or with dislocations from another pair.

Physically, the nucleation mechanism is related to the phase shifts induced by the transverse shear deformations. The reason for this is that the action of an external shear flow is cumulative in the phase θ of the complex envelope field $\psi = |\psi|e^{i\theta}$ and its gradients $\nabla\theta$, similar to the effect of the self-induced mean flow [92]. The shear flow advects the underlying stripe pattern together with its defects. This motion

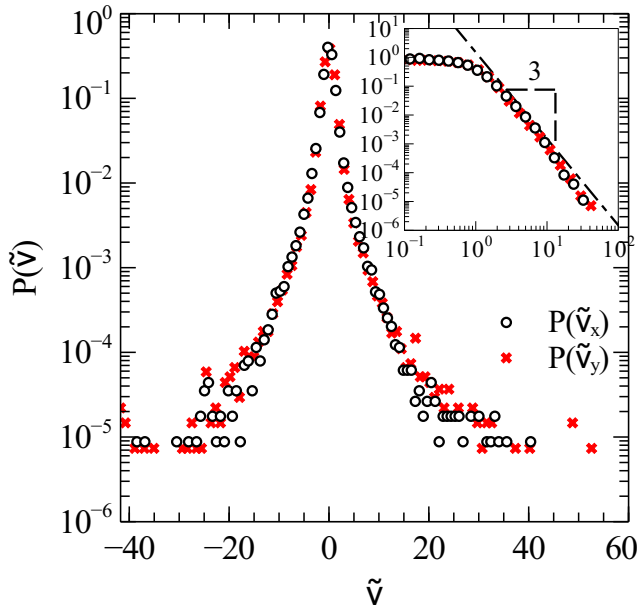


FIGURE 3.8: The time independent scaling function of the probability distribution corresponding to climb (crosses), respectively glide (open circles) velocities during the phase ordering in the absense of external field. The rescaled velocity variable is given as $\tilde{v}_j \equiv v_j/\langle|v_j|\rangle$ for $j = x, y$. In the inset figure it is shown the log-log plot of the PDF calculated with logarithmic binning.

induces small undulations along the stripes which build up stresses and create distortions in the pattern. These distortions, which are localized into transverse "shear bands", grow with time up to the point where they locally tear apart the stripes releasing pairs of defects. The shear flow affects both the isolated motion of defects and, more importantly, the interaction between defects. The effect of the large-scale flow on the defect interactions becomes stronger where the defect motion is slower [92].

3.4.2 Velocity statistics

In the absence of shear flow, the motion of dislocations is symmetric. Although the mean velocity in absolute value is non-zero and related to the mutual interaction forces, the velocity of dislocations averages out to zero. This is equivalent to the symmetric probability distributions of the velocity fluctuations as shown in Fig. (3.8). The actual distribution is time dependent due to quench dynamics by annihilations. However, using the dynamical scaling behavior of the probability distribution we can remove the time dependence by effectively rescaling the dislocation velocity by the ensemble average of the absolute velocity at a given time, i.e. $\tilde{v}_i \equiv v_i/\langle|v_i(t)|\rangle$, with $i = x, y$ for the two components. The distribution of these rescaled velocities corresponds to the scaling function of the time dependent probability distribution as discussed previously

in the context of vortex dynamics. From Fig. (3.8), we notice that the probability distributions of the climb and glide velocities retain a similar form that is characterized by a long tail with a -3 power law as in the relaxational dynamics of vortices. This is consistent with previous numerical studies of the velocity statistics of defects in the anisotropic Swift-Hohenberg dynamics from Ref. [59].

At a finite shear rate and in the late stage of statistically stationary regime, the motion of dislocations is influenced by the imposed flow. Dislocations of opposite topological charges tend to move in opposite directions, with an asymmetry in their mean transverse motion that is related to the shear rate, i.e. $\dot{\gamma} \sim (\langle v_x^+ \rangle - \langle v_x^- \rangle)$, while the climb motion remains almost symmetric. The velocity probability distribution for positive, respectively negative dislocations becomes the same when we rescale their corresponding absolute velocities as $\tilde{v}^s \equiv |v^s|/\langle |v^s| \rangle$, where $s = \pm$, so that $P^+(\tilde{v}^+) = P^-(\tilde{v}^-)$.

In Fig. (3.9), we plot the probability distribution functions of the rescaled gliding velocities and climbing velocities in the steady state regime. For the transverse motion, the small scale velocity fluctuations are normally distributed around the mean flow. Large fluctuations against the flow are due to events where pairs of dislocations of opposite charge, gliding opposite to their drift flow, are attracting and annihilating. These events contribute to the long left tail for $P^+(v_x)$ and right tail for $P^-(v_x)$. The statistics of small climb velocities are also influenced by the imposed shear flow and given by the slight asymmetry in the $P^s(v_y)$. However, the large fluctuations in the longitudinal motion are due to pair interactions prior to annihilations. These fluctuations are captured by the long tails in the $P(v_y)$ distribution and they seem to follow the inverse cubic law, but with less accuracy than in the relaxational case. Typically, a nucleation event leads to a burst in the local density of dislocations which will annihilate subsequently by the fast climbing motions. However, these large velocity events occur on a longer timescale than during relaxational dynamics, so that the system needs to be followed longer in the steady state. This is computationally challenging, because of the prior transient acceleration period whose length increases as the driving force is decreased.

3.5 Conclusions

In summary, our numerical simulations suggest that the velocity statistics of codimensional-two defects exhibits a dynamical scale invariance with a scaling function that has a universal inverse cubic tail when the defect dynamics is driven by mutual pair interactions leading to annihilations. This is valid both for point defects in 2D and defect filaments in 3D during phase ordering kinetics. Finite size effects of the defect core

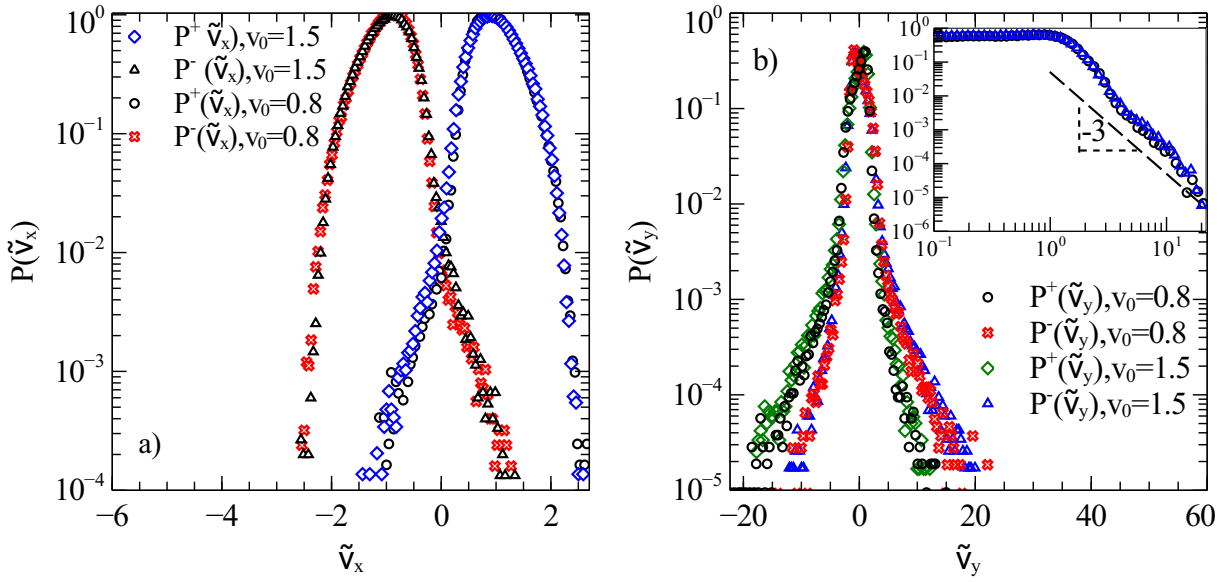


FIGURE 3.9: (a) PDF of the gliding velocity for positive (open diamonds and circles) and negative (open triangles and crosses) dislocations in the statistically stationary state at a different shear rates. (b) Probability distribution function of the climb velocity for positive (open circles and diamonds) and negative (open crosses and triangles) dislocations in the statistically stationary state. In both cases, the rescaled velocity variable is $\tilde{v}_j \equiv v_j/\langle|v_j|\rangle$ for $j = x, y$. Inset figure shows the log-log plot of the tail PDF with logarithmic binning.

introduce a Gaussian cut-off to the v^{-3} scaling. A similar statistical behavior is observed in the velocity of dislocations in anisotropic stripe patterns during phase ordering. Although the motion is highly anisotropic and dominated by the gliding of dislocations, the distributions of the climb and glide velocity fluctuations exhibit the same algebraic tail when the motion is driven by the local interactions between dislocations. During non-relaxational dynamics assisted by an external transverse shear flow, small velocity fluctuations are influenced by the mean flow, whereas the asymptotically large fluctuations are still due to pairwise interactions. In statistically stationary dynamics, anisotropic defect motion is manifested also in anisotropic statistics of the glide and climb velocities. In this study, we have neglected the effect of a self-induced mean flow in the defect dynamics. This effect is however important to capture the spatio-temporal chaotic dynamics as seen from experiments. It is thus interesting to study in future investigations the statistical properties of interacting defects when the combined effect of a large scale flow and a self-induced mean flow is taken into account.

Part II

Environmental and Evolutionary Microbiology

Chapter 4

On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys

Pyrosequencing platforms have been widely used in 16S rRNA deep sequencing of organisms sampled from environmental surveys. Despite the massive number of reads generated by these platforms, the reads only cover short regions of the gene, and the use of these short reads has recently been called into question for phylogeny-based and diversity analyses. We explore the limits of the use of short reads by quantifying the loss of information, and its effect on phylogeny. Using available nearly-full-length reads from published clone libraries and databases, and simulated short reads created from these reads, we show that for selected regions of the gene, short reads contain a surprisingly high amount of biological information, making them suitable to resolve an approximate phylogeny. In particular, we find that the V6 region is significantly poorer than the V1-V3 region in its representation of phylogenetic relationships. We conclude that the use of short reads, combined with a careful choice of the gene region used, and a thorough alignment procedure, can yield phylogenetic information comparable to that obtained from nearly-full-length 16S rRNA reads.

4.1 Introduction

Advances in sequencing technology allow researchers to generate massive libraries of biological information. In particular, high-throughput sequencing methods [93] are becoming widely used to analyze microbial communities [94–99]. One appealing aspect of recent advances in technology has been the deep environmental surveys of the microbial composition from a wide variety of environments, ranging from ocean [94, 100], to soil [101], to mammal guts [98]. In addition, 16S hypervariable tag sequencing has exposed the existence of a so-called "rare biosphere,"[95] whose contribution to, and impact in, the microbial environment are only now beginning to be observable, quantified and appreciated [99]. The potential significance of a previously unsuspected biosphere, rich in diversity but low in abundance, is that it may offer a major clue as to the response of ecosystems to change, and may well control their ability to adapt, by providing a large reservoir of genetic novelty to be tapped.

Despite this promise, the technology is still in its infancy. In particular, the reads generated by hypervariable tag pyrosequencing are short, spanning only hundreds of nucleotides. In the case of 16S rRNA, this has forced researchers to focus their studies on partial regions, usually including one or more of its hypervariable regions in an effort to capture the maximum possible amount of useful biological information. Naturally, there have been studies that compare the information obtained from these short reads with that obtained from nearly-full-length reads of SSU rRNA, quantifying the loss of information, dependence on the region of 16S rRNA being studied, and other possible biases [102–105]. In particular, the effects of the use of short reads in taxonomic assignments and ecological diversity indices have been documented. These studies make recommendations on which regions of the SSU rRNA are better suited to minimize the artifacts based on the observations of their studies, yet their recommendations are not fully consistent with each other, underlining some of the many complexities of the problem.

As pyrosequencing technology is maturing, the systematic artifacts that were present in earlier data sets have become less of an issue [95]. These artifacts were by no means minor in terms of their biological impact [106–108]. For example, point errors present in reads and artificial replicates lead to spurious operational taxonomic unit (OTU) counts and overestimation of abundances in the OTUs [106–108]. Fortunately, both of these artifacts can be easily removed with careful preprocessing [106, 108] and accurate alignment [15, 16] of the libraries. As these artifacts are being removed, we can grow more confident in pyrosequencing

data and focus on the challenges imposed by the intrinsic information loss present in these data sets.

The purpose of this chapter is to quantify the amount of phylogenetic information contained in short reads. In order to do this, we estimate the correlation between the phylogenetic information obtained using synthetic short reads, to that from nearly-full-length reads. To this end, we constructed an artificial clone library using 2000 nearly-full-length bacterial SSU rRNA sequences, randomly selected from the Greengenes [18] 16S database, retrieved August 2009. The sequences in the library were trimmed in length to simulate data obtained using pyrosequencing, creating additional libraries. The libraries were then used to construct Maximum Likelihood (ML) phylogenetic trees. To quantify how much information is preserved in the trees made with short reads, a branch-length-based pairwise distance metric [109, 110], supplemented with Robinson-Foulds [19] and weighted Robinson-Foulds [111] metrics, was used to correlate [110, 112] and compare the structures between the different trees. We show that two different inferences of a phylogenetic tree using the same short read library can show marked discrepancies due to the stochastic nature of maximum likelihood-based tree reconstruction methods the nature of the region being studied. We show that two different tree searches using the same short read library can show marked discrepancies due to the nature of the region being studied, given the randomized starting trees used by RAxML to perform such searches.

Then we show that, while a significant amount of information is preserved in the short read-based trees, the actual amount of information preserved seems to be not only a function of the length of the read, but also a function of the region sequenced. Our results indicate that the V1-V3 hypervariable region is a good estimator of phylogenetic information, and would be the preferred target for pyrosequencing assays of large communities, such as in large-scale environmental metagenomic surveys.

4.2 Materials and Methods

4.2.1 Selection of sequence sample and alignment

The single data set used in this study consisted of 2017 bacterial, nearly-full-length 16S rRNA sequences randomly selected from the Greengenes database [18], as of August 2009. The reason for choosing 2017 sequences as the sample size of the database comes from the desire to perform a realistic comparison in the light of the sizes of existing read libraries obtained for nearly-full-length 16S sequences (for example,

Tree pair	PC	RF	WRF
NM(1) vs. NM(2)	0.97	1042	179.50
NM(1) vs. LM(1)	0.93	1934	512.35
NM(1) vs. LM(2)	0.94	1900	495.39
NM(2) vs. LM(1)	0.94	1910	503.73
NM(2) vs. LM(2)	0.95	1864	484.81
LM(1) vs. LM(1)	0.96	1304	183.44

TABLE 4.1: Pearson correlation (PC), Robinson-Foulds (RF) and Weighted Robinson-Foulds (WRF) metrics for a non-masked (NM) and Lane-masked versions of a test alignment. Two tree searches were made from each alignment.

using Sanger sequencing). Although pyrosequencing is now able to create libraries with sizes in the order of millions of reads, a comparison study is certainly not realistic due to the lack of full-length libraries of that size in studies, and also bumping the separate problem of creating a phylogeny for such a big number of reads. Thus, we chose to limit ourselves to a simple case and small size, which is relevant to already published studies,

It is customary to perform Lane-masking [113] of 16S rRNA aqlignments when constructing Neighbor-Joining trees. In our work, we use the more accurate Maximum Likelihood (ML) algorithms. No masking of the alignment is needed, because a Maximum Likelihood approach would place little weight on extremely variable regions. To demonstrate this point, a Lane-masked version of an almost-full-length 16S rRNA alignment was tested using the standard Robinson-Foulds (RF) and Weighted Robinson-Foulds (WRF) metrics. The RF metric measures the the number of splits present in one tree that are not present in the other tree, that is, the symmetric difference of the two sets of splits. The WRF metric has the extra feature of weighting the splits by their support value. This alignment was used in two tree searches, and the corresponding results compared using these metrics. As shown in Table 4.1, the trees from the Lane-masked alignment show a RF distance of 1304, whereas the trees from the non-masked alignment show a RF distance of 1042. In the case of the WRF distance, the Lane-masked trees show a distance of 179.50, and the non-masked version shows a distance of 183.44. Although there is greater variation in the Lane-masked trees found by ML, these differences have low support values, as evidenced by the very similar WRF distances for both the Lane-masked and non-masked trees. This shows that Lane masking creates a measurable deterioration of the quality of the tree from a topological point of view, and therefore would create an uncontrolled artifact in our topologically-focused analysis. For these reasons we do not use Lane masking.

Tree pair	PC	RF	WRF
GG(1) vs. GG(2)	0.98	936	157.11
GG(1) vs. S(1)	0.95	1766	648.35
GG(1) vs. S(2)	0.92	1782	646.06
GG(2) vs. S(1)	0.96	1820	663.00
GG(2) vs. S(2)	0.92	1780	644.58
S(1) vs. S(2)	0.96	890	193.67

TABLE 4.2: Greengenes (GG) and SILVA (S) alignment templates compared using the Pearson Correlation (PC), Robinson-Foulds (RF) and Weighted Robinson-Foulds (WRF) metrics.

4.2.2 Influence of sequence alignment on tree metrics

The sequences extracted for this study were obtained from the Greengenes database, and as such those sequences are profile-aligned using NAST to Greengenes own 16S rRNA alignment template. If, for example, we profile-align the same sequences to a different template, such as the SILVA bacterial template, or an altogether different method, such as the Infernal aligner present in the Ribosomal Database Project's 16S pipeline, it is reasonable to expect differences between the obtained phylogenies. In this study we are interested in the relative differences between phylogenies processed using the same pipeline, so a valid question is, what differences can we expect in the tree comparison metrics if we use different alignment strategies?

To this end, we set up a simple control test to measure the differences between two different alignment templates, the Greengenes template and the SILVA template, starting from the same original data. We expect trees constructed using different alignments of the same library to be somewhat different, of course, but it is essential that trees resulting from searches performed on the same alignment should be more similar to each other (for full-length 16S trees) than to trees from a different alignment.

To see if this is the case, we re-aligned the full-length library to the SILVA bacterial profile using Mothur's NAST and compared the resulting trees to each other and to the trees from the "unaligned" library. The results are shown in Table 4.2. The trees were compared using Pearson Correlation, Robinson-Foulds (RF) and weighted RF (WRF) metrics. All three distance metrics between different alignments are greater than those for trees from the same alignment. This analysis shows that indeed there is consistency between trees made from the same alignment, and thus re-alignment is not necessary.

Although the RF scores of the SILVA and Greengenes alignments are comparable, the WRF scores are significantly different, presumably reflecting the presence of subtrees in the SILVA trees with higher support values than the Greengenes trees. This may also be reflected in the slight difference in the Pearson

correlation. In summary, re-alignment is not necessary for our analysis, and we proceed just using the original Greengenes alignment.

4.2.3 Trimming of sequences to create the libraries of simulated short reads

The trimming procedure necessary to create the artificial libraries was performed after the alignment of the source sequences was completed, instead of arbitrarily set the lengths before alignment, which would indiscriminately remove some information give the uneven starting and ending points for the reads.The reason for doing this is to maximize the information content of the reads, and also to use the standard starting and ending points for the studied regions, which are defined by the primers used at sequencing time. Also, for the almost-full-length library, the endpoints were trimmed in such a way to maximize the overlap between all reads.

The selected sequences were imported into the alignment manipulation program Jalview [114]. The sequences were then trimmed to the lengths expected for pyrosequencing reads coming from the 454 Life Sciences Genome Sequencers GS 20, GS FLX Standard and GS FLX Titanium platforms, making sure they contain the hypervariable regions of interest.

4.2.4 Construction of phylogenetic trees

To construct the maximum likelihood phylogenetic trees from the sequences in the libraries, we used RAxML [115] version 7.0.4, multithreaded using the Pthreads library [116], using the rapid bootstrap [117] option and the CAT model of rate heterogeneity [118]. The trees constructed using the nearly-full-length sequences were created from 300 BS replicates, and the ones created from the simulated short reads were created using 1000 BS replicates. We performed two tree searches for each library in the set using different seeds for RAxML's random number generator, which then we used to calculate pairwise distances.

The actual command line used for the tree searches reads as:

```
raxmlHPC-PTHREADS -T <threads> -f a -m GTRGAMMA -N <replicates>
                  -x <seed1> -p <seed2> -s <alignment> -n <name>
```

where "threads" is the number of threads per computer node, "replicates" is the number of BS replicates in the tree search, "seed1" and "seed2" are integer numbers used to seed RAxML's random number generators, "alignment" is the alignment file name, and "name" is a name to identify the output file.

4.2.5 Distance and correlation calculations

To compare the different trees, we used a definition of pairwise distance which depends on the structure of the tree. The pairwise distance between two sequences is defined as the sum of the branch lengths of the shortest path connecting the leaves representing the sequences in the tree [109]. For each tree we calculated the pairwise distances between all possible pairs of sequences, with branch-lengths calculated under GAMMA, thus creating a patristic distance matrix for a particular tree.

Then we calculate the Pearson correlation for all possible pairs of patristic distance matrices. For that, we create tuples from the corresponding elements in a pair of matrices, and the we calculated the Pearson correlation R^2 for this set of tuples.

We also supplemented this metric with standard metrics for tree comparison, namely the Robinson-Foulds (RF) metric [19] and a weighted Robinson-Foulds (WRF) metric [111]. The RF distance between two trees is the number of splits present in one tree that are not present in the other tree, that is, the symmetric difference of the two sets of splits. The WRF distance differs from the unweighted RF distance in that it assigns a weight to each of the splits present in the symmetric difference, and the actual value is then the sum of these weights. To calculate these distances between all trees created we used RAxML version 7.2.6, which uses the support values of the splits as the weight for WRF distances.

4.2.6 Plotting of distance data

As a data reduction step, we took the tuples created from the two distance matrices being compared, and proceeded to do a binning step. This step consisted in performing an average over intervals of size 0.01 distance units, and also obtaining the standard deviation for each interval. After normalizing the maximum distance to 1.0, and rescaling the standard deviations accordingly, the data sets were now ready for plotting.

4.3 Results

Since we aimed to compare phylogenetic information present in different regions of the 16S rRNA, we use a metric that can appropriately compare phylogenetic trees created from the different regions. Our metric measures the distance between a pair of sequences in the tree, and compares it to the corresponding distance in the other tree. The actual distance is computed as the length of the shortest path in the branches of the

tree that goes from one member of the pair to the other one [109]. This distance accounts for the different branch lengths calculated during the tree construction process. For a whole tree, this distance is computed for all possible pairs of sequences in the trees and it is stored as a distance matrix.

Now, to compare two trees, we create a tuple of the distances of corresponding pairs in both trees, and then we calculate the Pearson correlation R^2 for the set of distances [112]. We chose this method because it provides a balance between the computational expense of calculating more detailed comparison metrics, and the oversimplification of comparing single numbers coming from each tree, numbers that do not necessarily provide meaningful information, unless supplemented by extra details, such as the distribution of possible distances between randomly structured trees, which itself is something still very time-consuming to calculate.

We expect, for very similar trees, that the correlation R^2 will be very close to 1.0. This is a consequence of how close or far we expect different pairs of reads to be in different trees. In an ideal case, reads that are found to be close together in one tree are also expected to be found close in the other tree. Likewise, if two reads are found to be far apart in one tree, they are expected to be distant in the other tree. Thus, if we plot the distances found in one tree as a function of the distances found in a second tree, we would obtain a straight line in the case of identical trees, with $R^2 = 1.0$. For trees that are slightly different, the points will be scattered around this straight line. When the correlation is very poor, this straight line behavior would be just a weak trend buried in a jungle of wildly scattered points. That said, we do not expect a situation where we observe very good, non-linear correlation, which can give very low values of R^2. We will call these graphs Sequence Correlation Plots.

As a first calculation, we applied this metric to different trees created from the same region. For the nearly-full-length case, this measurement would yield the minimum possible uncertainty in the structure of the resulting trees, resulting from the ML procedure we used. Thus, we construct an upper limit on the quality of our comparison methodology. From the resulting value of R^2 and the graph related to the comparison, we can base subsequent explorations.

The short reads are between 120 and 400 base pairs long, or approximately 15% and 30% of the full length of the 16S gene, and it is reasonable to expect a loss of phylogenetic information when using these reads. Now, when this distance metric is applied to trees created from the simulated short reads, we expect more deviations from the straight line behavior of almost identical trees. This is because the ML trees

created from shorter reads are harder to resolve –less sequence information leaves greater ambiguities to be resolved. The ML problem in these cases requires a more intense search of the solution space in order to converge to a solution, which itself is one of many equally good approximations to the "real" solution. This suggests an interesting possibility: if it is harder to find a solution to the ML problem, that means the sequences would contain less phylogenetic information. This implies that the quality of a ML tree for fixed computational effort could be used as a proxy for measuring phylogenetic information content in the sequences used to create the trees.

As a way to supplement the patristic correlation metric, and also to gain further insight on the difference between the trees, we also compared the trees using a Robinson-Foulds (RF) metric [19], and a weighted Robinson-Foulds (WRF) metric [111]. The RF metric is the count of the splits present in one tree that are not present in the other. In other words, it is the symmetric difference of the sets of splits of the trees being compared. The WRF metric, on the other hand, multiplies each split count by a certain number. In our case, each split is weighted by the support value of said split, where the support value goes from 0 for no support, to 1.0 for full support for the split. Besides providing another measure of similarity between trees, using both RF and WRF metrics provide us some insight on the nature of the differences between the trees. If the RF distance is much larger than the WRF distance, we can infer that the differences between trees occur mostly on low-support subtrees, whereas if the WRF distance is closer to the RF distance, then the differences are mostly due to rearrangements of high-support subtrees [119].

From this kind of comparison, we can gain some insight on two specific questions we have about the short reads: (i) how much information is contained in the hypervariable tag regions; and (ii) how the length of the read correlates with the phylogenetic information contained in the nearly complete gene. There has been previous work regarding the latter question, exploring simulated datasets to address the general question of length requirements for tree-reconstruction [120, 121], and although their concerns extend well beyond the question of short-reads and into large-scale phylogeny in general, their findings are certainly relevant to our case and add to our observation of the complexity of the short-reads situation.

In Figure 4.1, we show Sequence Correlation Plots (SCP) for the nearly-full-length (FL) tree, as well as for all the considered regions. Distances in the plot are normalized to a maximum value of 1.0 and, for clarity, they are also binned. The bins have a width of 0.05 distance units before normalization. The error bars correspond to the standard deviation calculated during the binning process.

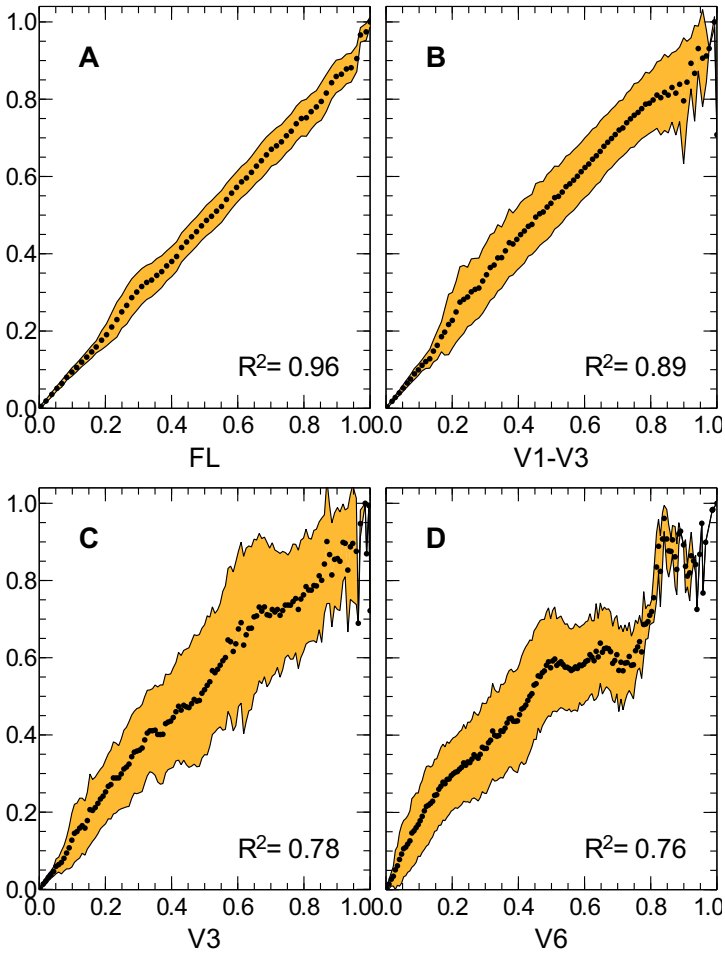


FIGURE 4.1: Correlations between tree searches performed on full length and partial reads. The sequence correlation plots show how similar are two ML tree searches for the different regions considered. The higher the correlations, the more similar the solutions to the ML problem. This measure can be used as a proxy for phylogenetic information content in the region used to construct the tree. The dots correspond to the average normalized distance in one tree as a function of the corresponding distance in the other tree, averaged over all distance pairs in a bin. The shaded area corresponds to the standard deviation for the points in each bin.

In Figure 4.1A, we show the comparison between two tree searches performed on the alignment of nearly-full-length reads. The correlation of these two trees is very high, and this comparison can be thought of as an approximate minimum of the possible uncertainty in determining the phylogeny based on 16S rRNA, using RAxML as the treeing tool, with the NAST-based alignments. There are differences even in the nearly-full-length trees because of the approximate, probabilistic nature of the solutions found for the ML problem as solved by RAxML. Even in this "benchmark" case, there are a couple of notable features in this plot. First, the error bars are very small in the lower 20% of the pairwise distances in the plot,

meaning that the reads that were found to be close together in one tree search stayed in similar positions in the other search. As the pairwise distance is increased, there is an increase in the size of the error bars, yet the error bars stay relatively constant up to the maximum distance in the trees. This indicates that the relationship between distantly related pairs does not significantly deviate from the average, indicating that the relationships between pairs are preserved on average.

In Figure 4.1B, we see a very similar situation for the reads encompassing hypervariable regions V1 to V3 (V1-V3), a library that can be experimentally obtained using 454 GS FLX Titanium platforms. The rather surprising observation is that the lower 15% of the pairwise distances are very well preserved between tree searches on the same alignment, the error bars being comparable to those present in Figure 4.1A. Beyond this 15%, the error bars grow substantially, but their magnitude remains more or less constant throughout the set of pairwise distances, signaling some loss of information compared to the FL trees, yet, at the same time, still being able to resolve the long distance relationships between reads for a majority of the pairs. Interestingly, the Pearson correlation in this case ($R^2 = 0.89$) is very similar to that obtained when comparing the V1-V3 tree with the nearly-full-length reads tree (see Figure 4.2). Overall, when compared to the nearly-full-length situation, the V1-V3 has produced very consistent trees, as indicated not only by visually comparing the graphs, but also by the high value of R^2.

The situation for the other two regions appears very different. There is poor consistency in the structures of the different tree searches, and the error bars tend to grow as the distance increases. In the V6 region graph, there is a major change in behavior at very large distances, signaling substantial differences in the tree structures.

Figures 4.1C and 4.1D show the resulting plots for the shortest reads analyzed. Beyond the very closely related reads, the error bars keep growing as the pairwise distances grow, in contrast to the FL and V1-V3 reads. This likely indicates an inability to consistently resolve the relationship between distant reads across tree searches. In particular, when focusing on the two trees we constructed using the V6 region (Figure 4.1D), there are major differences between the distantly related taxa between both trees. This, despite the fact that the trees constructed using the short reads were calculated using more BS replicates during the construction process, and it is likely that adding more BS replicates to this process will not significantly change the outcome.

We now apply the same comparison metric to trees made from different regions to explore cross-

correlation between different regions of the 16S gene. Unlike the previous case in which we used the correlation metric as a proxy to estimate the intrinsic phylogenetic information content of a given region, this cross-correlation between different regions can be used to estimate the phylogenetic information content from a region, relative to the information present in another region.

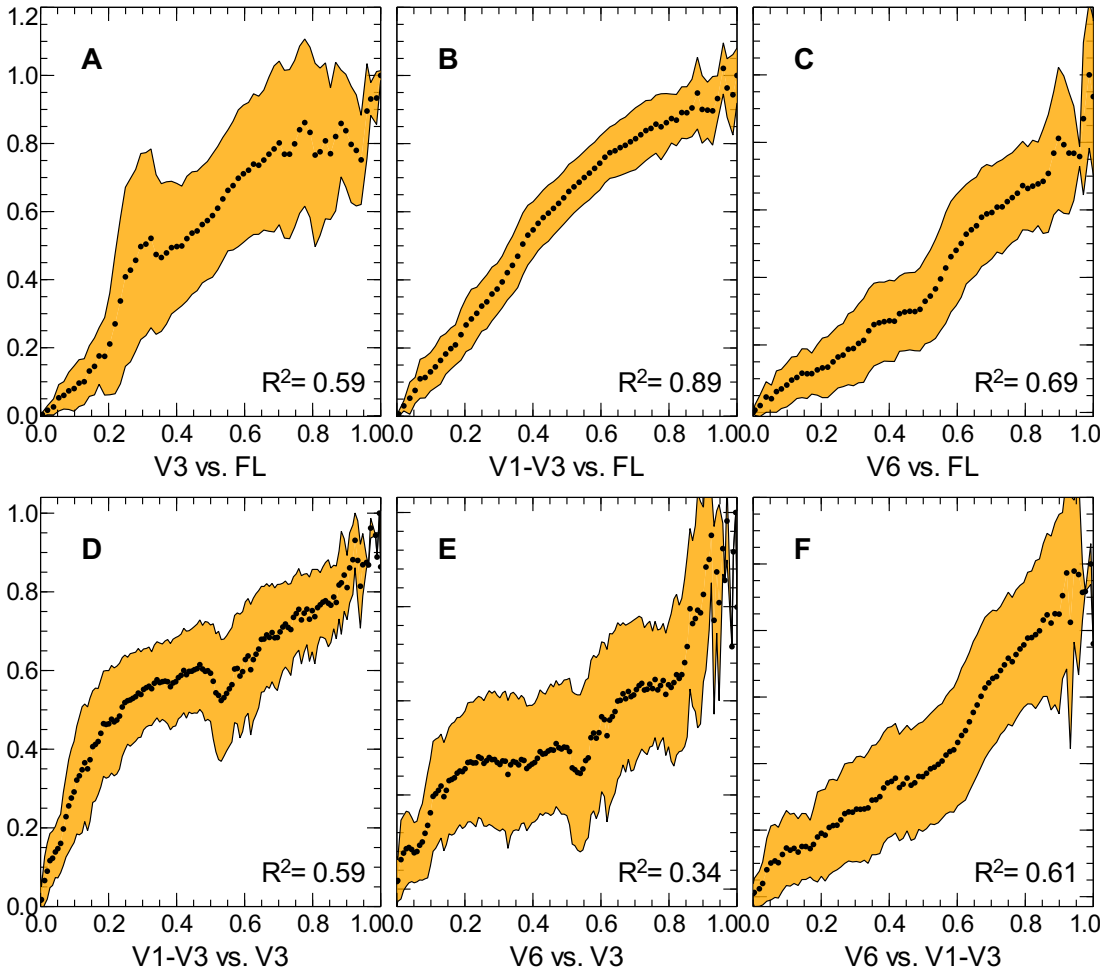


FIGURE 4.2: Correlations between trees constructed using different regions. The top row shows pairwise distance correlations between trees from the different regions and a full length tree. The bottom row shows the correlations between all the trees made from the different regions. These graphs illustrate the possible variations in tree correlation, ranging from very good (B) to poor (E). The dots correspond to the average normalized distance in one tree as a function of the corresponding distance in the other tree, averaged over all distance pairs in a bin. The shaded area corresponds to the standard deviation for the points in each bin.

Figure 4.2 shows the normalized pairwise distances between the reads in one tree, as a function of the corresponding distance of a second tree. The top row contains all the comparisons against the nearly-full-length reads, where the V1-V3 tree shows good agreement with the FL tree. Other comparisons for this

same region show similarly high correlations (see Table 4.3). A different picture can be observed for V3 (Figure 4.2A) and V6 (Figure 4.2C), where correlations are not as good. The values of the correlations also vary with tree searches, such that the better match between a short read and the nearly-full-length tree is dependent on the found trees, signaling a rather important loss of phylogenetic information.

The bottom row shows the comparison between the different simulated short read libraries. As expected, comparisons with the V1-V3 region reads give a better correlation, comparable even with the correlations with nearly-full-length reads. But the comparison between the V3 and V6 is not as good, highlighting substantial differences between the structures of the trees created using these regions.

The Robinson-Foulds (RF) and weighted Robinson-Foulds (WRF) distances calculated for all tree pairs, besides complementing the data obtained with the correlations and giving us more insight on the differences between the trees, should also be consistent with the trends shown in the figures. Table 4.3 contains the Pearson correlations (PC), RF distances and WRF distances for all tree pairs analyzed in this study. The number in parentheses indicates which of the tree searches for that particular region is being compared. The purpose of the table is to show all the range of R^2 values present in the dataset, and also compare them with the corresponding RF and WRF distances. The first point we want to show is the range of differences for trees created from the same alignment. The PC values drop from the FL trees down to the V3/V6 case, as the RF distances increase, and the WRF distances only show a slight increase. We can interpret these differences as an increasing discrepancies between topologies and branch lengths as the size of the regions being used to construct the trees become shorter, but given the WRF distances most of the differences seem to be concentrated on the low-support subtrees. The next trend we notice is that the R^2 values between the V1-V3 trees and the FL trees are consistently high, ranging from 0.85 to 0.89. Their corresponding RF distances are very similar, ranging from 2700 to 2742, and their respective WRF distances also follow this pattern. These numbers show that, while there is quantifiable phylogenetic information loss, the tree structure is rather well resolved and consistent across tree searches. The next point we want to highlight is the fluctuation of the R^2 values for the trees found for V3 and V6, when correlated to the FL trees. From these numbers alone we cannot reliably tell if one of these regions is more suitable than the other. However, by looking at the RF and WRF values, we can see that the topological differences between the trees are due to rearrangements of high-support subtrees. The fact that the RF and WRF distances don't fluctuate as much compared to the PC values, shows that, while the differences in subtree structure might be relatively

Tree Pair	PC	RF	WRF	Tree Pair	PC	RF	WRF
FL(1) vs. FL(2)	0.96	1012	150.98	V1-V3(1) vs. V3(1)	0.59	3302	746.68
FL(1) vs. V1-V3(1)	0.89	2700	861.52	V1-V3(1) vs. V3(2)	0.64	3336	750.52
FL(1) vs. V1-V3(2)	0.85	2724	871.21	V1-V3(1) vs. V6(1)	0.61	3634	942.84
FL(1) vs. V3(1)	0.59	3310	980.40	V1-V3(1) vs. V6(2)	0.53	3600	946.27
FL(1) vs. V3(2)	0.68	3308	969.56	V1-V3(2) vs. V3(1)	0.63	3306	746.42
FL(1) vs. V6(1)	0.69	3472	1083.62	V1-V3(2) vs. V3(2)	0.63	3322	744.12
FL(1) vs. V6(2)	0.60	3444	1083.58	V1-V3(2) vs. V6(1)	0.55	3638	943.23
FL(2) vs. V1-V3(1)	0.86	2702	871.29	V1-V3(2) vs. V6(2)	0.46	3596	946.51
FL(2) vs. V1-V3(2)	0.88	2742	886.41	V3(1) vs. V3(2)	0.78	2560	159.82
FL(2) vs. V3(1)	0.65	3300	979.69	V3(1) vs. V6(1)	0.34	3716	675.30
FL(2) vs. V3(2)	0.70	3292	969.87	V3(1) vs. V6(2)	0.33	3710	687.56
FL(2) vs. V6(1)	0.65	3476	1093.31	V3(2) vs. V6(1)	0.47	3704	666.35
FL(2) vs. V6(2)	0.58	3450	1091.76	V3(2) vs. V6(2)	0.42	3700	680.00
V1-V3(1) vs. V1-V3(2)	0.89	1578	171.42	V6(1) vs. V6(2)	0.76	2772	195.53

TABLE 4.3: Correlations and distances between all trees studied. The table shows Pearson correlations (PC) between all the trees used in this study, as well as their corresponding Robinson-Foulds (RF) and weighted Robinson-Foulds (WRF) distances. The number in parentheses represents the tree search used when multiple trees were made from the same library. We see that the PC values for the longest short read libraries (V1-V3) are consistently high, and the corresponding values for the V3 and V6 regions show a high degree of fluctuation. Also of interest is the increase of RF/WRF when moving from V1-V3 towards V6, noting that the differences between trees start involving high-support subtrees, specially in the V6 case.

comparable, the branch length differences have a measurable effect, signaling potentially damaging loss of phylogenetic information for these regions. The final point is that the magnitude of the values of R^2 for the V3 and V6 trees, when correlated with the V1-V3 trees, are only slightly lower than for the FL case, the corresponding RF distances are similar, and their WRF distances are slightly lower, meaning that the topological differences are slightly more biased towards lower-support subtrees, giving another point of support for V1-V3 as being a good proxy for the FL reads.

Finally we can say something about the trees themselves. In Table 4.4 we see the tree lengths, defined as the sum of all branch lengths in the trees, and the logarithm of the likelihood (LogLk) value for the trees, as calculated using the original full-length alignment. The values of the tree lengths don't seem to follow a particular trend with respect to the read length. On the other hand, the LogLk values seem to get closer to zero as the read length decreases, meaning that the trees obtained are a progressively worse description of the data in the full-length alignments, when decreasing the read lengths, thus adding another layer of support to the observations about the discrepancies when using shorter reads.

Tree	Length	LogLk
FL(1)	189.38	-473754.97
FL(2)	188.15	-473750.26
V1-V3(1)	252.90	-494070.55
V1-V3(2)	254.76	-493119.36
V3(1)	172.54	-519286.99
V3(2)	187.78	-520317.05
V6(1)	132.72	-552733.64
V6(2)	136.26	-550485.72

TABLE 4.4: Parameters characterizing the trees. The table shows the tree length, defined as the sum of all branch lengths of the tree, and the logarithm of the likelihood for the tree (LogLk) using the full length alignment as the input data, as calculated during the tree search using a Maximum Likelihood method. The tree lengths don't seem to follow a trend with decreasing read length. The likelihood values, on the other hand, indicate that the trees made with short reads get worse at describing the relationships inferred from the full-length alignment as the reads become shorter.

4.4 Discussion

We have examined trees created from simulated short read libraries that were constructed using a random sample from the Greengenes database, comparing them using a Pearson Correlation of patristic distances between leaves of the trees, supplemented with Robinson-Foulds and weighted Robinson-Foulds distances. This comparison give us insight on the phylogenetic information content of the short reads, and it is a complement to other studies that quantify the pros and cons [106, 107] of pyrosequencing technology. This is specially relevant now, in the light of the huge influx of environmental data coming from big projects such as the ocean environmental studies [122] , Human Microbiome Project [123, 124], or studies on other more particular environments [125, 126] that would be difficult to accomplish if not for pyrosequencing.

We have also only considered focusing on a *de novo* approach to constructing phylogenetic trees, instead on doing a survey on all popular methods for reconstructing phylogenies, including insertion of reads into a pre-existing tree. Although the question of the reliability and comparison of large-scale phylogenies is certainly interesting, as evidenced by published studies on the subject (for example [103]) we wanted to focus on a specific problem of phylogenetic information loss, not doing a comparison of methods used in community analysis, which is beyond the scope of this chapter.

Our study identifies in detail the limitations of the short reads, from a phylogenetic information point of view, complementing other short read studies [103–105], which conclude that short reads less than 200 bp long show significant topological differences between tree searches, signaling phylogenetic information loss. From this observation, we can say that any conclusions derived using phylogeny-based tools (most

notably Unifrac [127]) that used very short reads and *de novo* phylogenies as their input data, have to be interpreted with some significant degree of uncertainty, independent of the region of 16S sequenced. Similar concerns have also been expressed elsewhere in the published literature [128].

On the other hand, based on the results shown here, the prospect looks much better when using appropriately chosen longer reads, which are already accessible using FLX Titanium pyrosequencing technology. These longer reads make it possible to extract phylogenetic information with high degree of reliability. The type of analysis we performed can be extended to other genes of interest, such as proteorhodopsins [129], which show a high degree of environmental correlation.

4.5 Conclusion

In this chapter, we generated synthetic short reads from complete 16S rRNA databases, and compared the complete phylogenetic trees with those obtained from the synthetic short reads. Our results show unequivocally that the different hypervariable regions are not equally suitable for this purpose, and that the V1-V3 region is the one that represents the best proxy for the complete 16S rRNA gene.

Chapter 5

Robust computational analysis of rRNA hypervariable tag datasets

Next-generation DNA sequencing is increasingly being utilized to probe microbial communities, such as gastrointestinal microbiomes, where it is important to be able to quantify measures of abundance and diversity. The fragmented nature of the 16S rRNA datasets obtained, coupled with their unprecedented size, has led to the recognition that the results of such analyses are potentially contaminated by a variety of artifacts, both experimental and computational. Here we quantify how multiple alignment and clustering errors contribute to overestimates of abundance and diversity, reflected by incorrect OTU assignment, corrupted phylogenies, inaccurate species diversity estimators, and rank abundance distribution functions. We show that straightforward procedural optimizations, combining preexisting tools, are effective in handling large (10^5 – 10^6) 16S rRNA datasets, and we describe metrics to measure the effectiveness and quality of the estimators obtained. We introduce two metrics to ascertain the quality of clustering of pyrosequenced rRNA data, and show that complete linkage clustering greatly outperforms other widely-used methods.

5.1 Author summary

Microbes constitute the majority of the living mass and genetic diversity on the Earth. They construct communities with elaborate patterns of abundance and diversity. Major current scientific interest is in quantifying these patterns in various microbial ecologies, from microbes living in oceans, to microbes in vertebrate

intestines and on human palms. To ascertain the diversity and abundance of microorganisms in a sample, experimenters sequence organismal tags. However, it is important to carefully analyze these data sets in order not to overestimate the number of species living in an environment. Our chapter describes how to make correct computational analyses of the next generation of deeply sequenced genome populations, where the unprecedented sample size and fragmented data set pose special problems to conventional approaches. As more and more sequence data are collected, it becomes critical to have robust, reliable and fast computational methods for analyzing these data. We have developed and made available online the software TORNADO (Taxon Organization from RNA Dataset Operations) to perform such computational analysis, and to quantitatively measure its quality.

5.2 Introduction

There is a long history of using environmental 16S rRNA [130] to estimate microbial diversity [131]. While early techniques relied on using clone libraries [132], next-generation high-throughput sequencing technology, such as pyrosequencing, directly generates vast libraries of sequences [133]. Next-generation high-throughput sequencers are capable of producing large datasets of more than a million reads from a single plate [134]. As the size of these datasets grow, the ability to computationally manage and characterize such data becomes a larger and more critical component of microbial ecology.

The goal of analyzing these sequences is to quantify the diversity and abundance distributions of organisms present in the environment. As pyrosequencing technology advances, our ability to measure microbial diversity increases. Already, this technique has been used to study the diversity of microbiomes from a variety of environments [135, 136], resulting in reports of a so-called "rare biosphere" of low-abundance organisms [95].

In order to assess the microbial diversity present in any dataset, the ability to appropriately measure the distance between different sequences and to reliably group them into operational taxonomic units (OTUs) is paramount. Typically, the abundance of OTUs is plotted in a rank abundance plot. These plots have been used as a gold standard for ecological population modeling for many decades[137]. In addition, the OTU groupings are utilized by other metrics in determining relative species compositions, microbial diversity, and community comparisons[138, 139]. While much effort and controversy has been focused on measurements of the quality of next-generation sequences[95, 103, 105, 108, 140], or the interpretation of pyrosequenc-

ing flowgrams [106], less attention has been given to computational analysis of pyrosequenced 16S rRNA data after quality processing, despite the large discrepancies in OTU numbers and diversity when different analysis methods are used [106–108, 141].

There are two major components to the analysis of OTU abundance. The first is multiple alignment of 16S rRNA or fragments of 16S rRNA. The second is clustering the sequences based on a distance metric. The purpose of this chapter is to provide a careful discussion of the computational analysis of alignment and clustering of pyrosequencing datasets, identifying sources of error, and appropriate ways to handle the data to mitigate these artifacts. In particular, we use the Calinski-Harabasz (CH) index [142] to compare the quality of clustering, finding unexpectedly large differences in the performance of different algorithms. We show that data analysis is surprisingly sensitive to even small errors in multiple alignment and clustering, but that with relatively little difficulty, these artifacts can be substantially mitigated using a judicious combination of preexisting tools, and others that we have made available on a web site (`http://tornado.igb.uiuc.edu`). Following these procedures results in robust characterization of microbial ecosystems.

5.2.1 Multiple Alignment: NAST and Infernal

Multiple alignment is the starting point of almost all analyses performed on microbiome sequences. Most phylogeny [115, 143], community distance estimates [139, 144], and abundance distributions [145–147] ultimately rely on input from a multiple alignment to compute sequence distances within a consistent alignment template.

The goal of multiple alignment is to align sequences according to their evolutionary relationships. In order for a multiple alignment to be meaningful in this context, all sequences in the multiple alignment must have a common origin. The various match, mismatch, and indel events then represent possible reconstructions of the evolution of those related sequences. In contrast to pairwise alignment, multiple alignment leverages conserved features of an entire gene family to obtain a broader evolutionary picture. This picture can then be fed into various algorithms such as maximum-likelihood phylogeny[113, 115, 143] in order to reconstruct the evolutionary relationships between the individual sequences.

The use of 16S rRNA sequences for discerning evolutionary relationships has a long history. The very first studies that organized the Bacteria according to their evolutionary relationships and resulted in the

discovery of the Archaea utilized this important ribosomal molecule as a molecular fossil [148] and it still remains the most widely used evolutionary marker in microbial ecology today [95, 98, 134]. As such, it is not surprising that a number of tools exist which are specifically tailored to 16S rRNA such as the NAST pipeline [15] or Ribosomal Database Project [16].

These specialized 16S rRNA alignment tools all incorporate information about the 16S rRNA secondary structure. The importance of the secondary structure is two-fold. First, the conservation of 16S rRNA sequences stems from the conserved structure. Second, unlike proteins which are built from up to 20 different amino acids, there are only 4 basic RNA bases. Randomly chosen RNA bases have a greater chance of aligning well with one another than randomly chosen protein sequences, making it more difficult to distinguish between evolutionary relationships and random matches. Secondary structure can, and should, be used to provide extra discriminatory power beyond that available from the one-dimensional sequence alone.

The NAST algorithm [15] tries to align new sequences against a precalculated multiple alignment template, and has been integrated into many commonly used 16S rRNA analysis tools such as Mothur [139] and GreenGenes[18]. Typically, this template is hand-curated to include the appropriate secondary structure considerations. In this chapter we will use the SILVA SEED SSURef database version 102 [149] as the template and refer to this alignment method as NAST+SILVA. The weaknesses of this method are that errors in the hand-curated multiple alignment propagate and that alignment against a fixed-size template necessitates the inclusion of purposeful misalignments. Overall, this results in alignments that are sometimes inconsistent with alignments based on secondary structure. An example of this is shown in the alignments in Figure 5.1b. By contrast, Infernal [17], which has been integrated into the Ribosomal Database Project 16S rRNA Pipeline [16], aligns sequences against a predefined structure. However, even among the well-conserved structures of 16S rRNA, there exist hypervariable regions which vary in their secondary structure from taxon to taxon. These regions cannot be aligned to a fixed structural template, and are left unaligned by Infernal (leading to a multiple alignment whose length is not fixed but may be different in different datasets). An example of this Infernal's alignment is shown in Figure 5.1a. It is important to note that while both methods align to a seed model of some sort, the practical difference is that RDP+Infernal does better in regions of strong secondary structure whereas NAST+SILVA does better in hypervariable regions.

To exploit this distinction, one can merge the best alignments from each tool by combining the hypervariable regions aligned using the NAST algorithm with the regions of strong secondary structure aligned

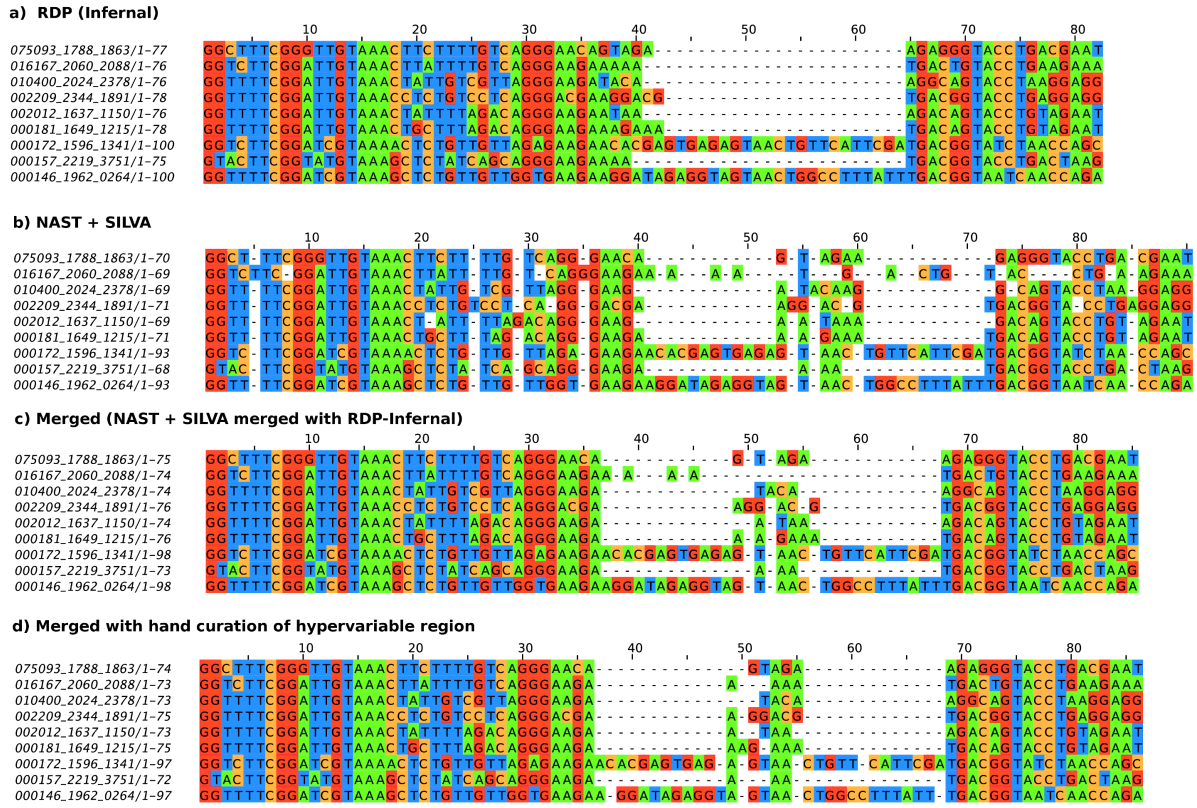


FIGURE 5.1: Snippets of 9 reads aligned using the 4 different methods described in this chapter. The 9 reads are of the V3 region of the 16S rRNA. (a) Sequences aligned via RDP [16] which uses the Infernal aligner [17]. Note that the hypervariable region is left unaligned (bases 36 through 64). (b) Sequences aligned via NAST [15] (as implemented by Mothur [139]) to the SILVA [149] database. Notice the inconsistencies in the alignment of the regions with strong secondary structure conservation (bases 5, 25, 29, 72 through 79, and 85). (c) Sequences aligned using the merge program in the tool we developed, TORNADO (`http://tornado.igb.uiuc.edu`). The merge process takes the unaligned, hypervariable parts of the sequence aligned by (a) and replaces them by the alignment in (b). (d) Sequences aligned like in (c), but with the final hand-curation step of the hypervariable regions.

by Infernal. An example of sequences aligned using the merging method is given in Figure 5.1c. One can also make adjustment to the multiple alignment by hand. Done properly, this can produce a better quality alignment than automated methods alone. An example of the merged alignment in which the hypervariable region was further hand-curated is given in Figure 5.1d. For a brief description of the process and the tools that we developed to perform the merging and hand-curation, please refer to the Materials and Methods section.

With the availability of a variety of tools that perform multiple sequence alignment, it is imperative to have a way to assess its quality. One way to do that is through maximum-likelihood (ML) phylogeny. ML phylogeny tries to identify the set of relationships with best likelihood value. Conversely, ML scores

can also be used to judge the likelihood of a multiple alignment reflecting sequence evolution. Indeed, similar measures have been used in the past [150] and tools such as SATÉ [151] already take advantage of this measure when iterating between multiple alignments and phylogeny to automate the search for the best alignment and tree. However, exploring enough multiple alignments for large datasets is prohibitively expensive and therefore remains impractical for now. Nonetheless, it is feasible to use the ML method in order to compare the quality of alignments by measuring their likelihood values, and we use this below to compare different alignment strategies.

5.2.2 Clustering Algorithms

Clustering algorithms, such as complete linkage [152], are essential for quantifying the diversity of microbial communities. The goal of clustering is to group sequences that are within some measure of evolutionary distance. Distances can be calculated using many different metrics such as percent sequence identity (PSI) or distance along the phylogenetic tree branches. Ideally, a clustering algorithm should identify the natural boundaries between the clusters without utilizing more clusters than necessary to account for the entire dataset. Ultimately, clustering should accurately reflect the underlying phylogenetic and taxonomic distribution of sequences.

Complete linkage clustering (as implemented by Mothur [139]) has become the most widely-used clustering algorithm in microbial ecology. It relies on input from a distance matrix that can be generated from the pairwise distances between sequences in a multiple alignment. When calculating sequence distances, it is important to clearly note how alignment gaps are dealt with. One can ignore the gaps (like Phylip DNADIST does [147]) or count them in a number of different ways [145]. Once pairwise distances are obtained, complete linkage operates by progressively merging smaller clusters into larger ones, as long as each element in a cluster is within a defined distance from other elements in the cluster[152].

The performance of linkage clustering algorithms, for example as implemented in Mothur, scales poorly (N^3, where N is the number of sequence reads) as the number of sequence reads generated per study increases. In the studies reported below, we reimplemented Mothur's clustering algorithm, achieving an improvement in computational complexity (scaling as $N^2 \log N$), better memory usage, and an overall speedup that is typically a factor of 5-10, leading to the ability to handle datasets with N up to about 30,000. Despite these improvements, it is understandable that heuristic, computationally efficient algorithms have been

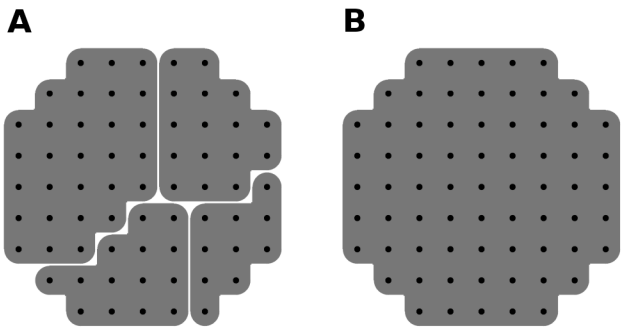


FIGURE 5.2: Calculation of clustering a set of points in a plane. (a) FastGroup's method. (b) complete linkage clustering. Both of these clusterings are performed with the same radius r equal to the radius of the set of points. FastGroup constructs 4 clusters whereas complete linkage finds 1.

developed, such as FastGroup [153] and ESPRIT[154].

FastGroup does not order clustering in any particular way, but instead chooses a sequence at random, grouping everything within a defined PSI distance of that sequence. As an example of how that is different from the complete linkage clustering employed by Mothur, consider the clustering of a scatter of points in two dimensions, as shown in Figure 5.2. The two-dimensional space is a very simple example of sequence space, with position in the space corresponding to the particular sequence of an organism. A set of points in this space, if sufficiently close to one another, represents a set of sequences that can be considered to be grouped into a single equivalence class—in other words, an OTU. The largest allowable distance between points in a single equivalence class corresponds to the sequence similarity required for sequences to be included in the same OTU (typically 97% is used).

When the FastGroup algorithm is used to group these sequences, with a radius equal to the radius of the circle of points, the number of clustered OTUs can vary, depending on the order of chosen cluster centers. One example of FastGroup's clustering is given in Figure 5.2a. On the other hand, complete linkage clustering with the same diameter correctly identifies the existence of 1 cluster (Figure 5.2b), by progressively merging clusters as long as they are within a cluster diameter (see Figure 5.3 for the progress of the complete linkage algorithm).

ESPRIT[154] goes one step further and does away with multiple alignment entirely and processes the clusters in two steps: the first relying on a k-mer heuristic, similar to that used in BLAST[155], in order to group closely related sequences under one representative sequence; the second relying on pairwise distances between representatives in order to determine the final clusters. Both FastGroup and ESPRIT differ from the

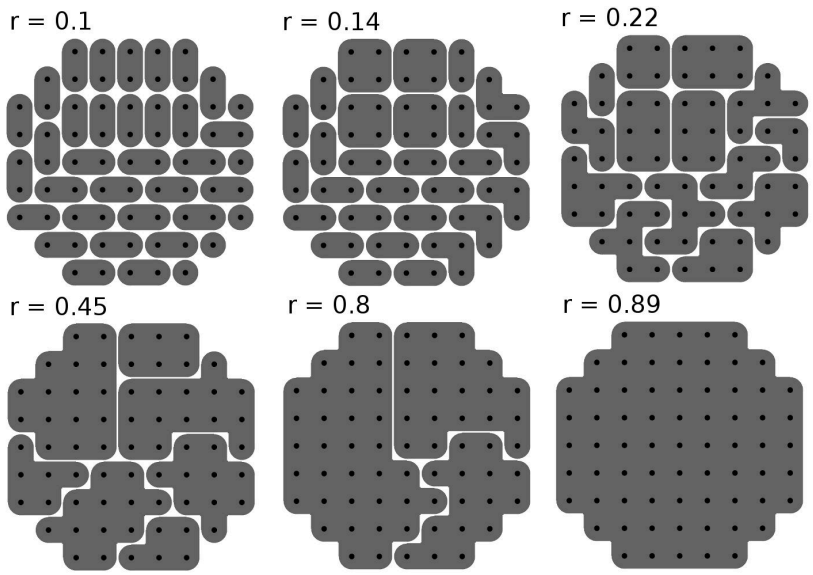


FIGURE 5.3: Illustration of the process of the complete linkage algorithm. Smaller clusters are progressively merged into larger ones as long as no two elements of a cluster are farther than r from each other.

more controlled calculations of the complete linkage algorithm, but at the same time promise less computationally intensive results. Before pursuing such alternatives, it is important to understand the differences between the results produced by each of these algorithms.

In other words, do the heuristic algorithms produce natural cluster borders and correct cluster compositions? A natural cluster should have a representative sequence that is near the center of the cluster, i.e. the representative sequence should be one that shares the most similarity to all other sequences in the cluster. Natural clusters should not partition the dataset into more groups than necessary. One way to quantify this goodness of clustering is via the Calinski-Harabasz (CH) index [142] that has been found to be the best in a comprehensive study of 30 different clustering quality indices [156]. In essence, the Calinski-Harabasz index is higher when the cluster centers are further away from each other (i.e. the clusters are better delineated from each other), and when the cluster radii are smaller (i.e. the clusters are tighter). The CH index is also correctly normalized so as to be comparable for different number of OTUs.

5.3 Results

In this work, we demonstrate that different methodologies can lead to very different estimates of OTU abundances. We characterize these differences and deconstruct their two primary sources: multiple alignment and the clustering method used. We measure the performance of both components of this process, restricting

ourselves to 16S rRNA based techniques. We also provide metrics to quantitatively evaluate the effectiveness of algorithms used. Our analysis includes an examination of the robustness of these algorithms on real biological data. We perform our analysis on a dataset of $22,911$ bacterial 16S rRNA sequences (V3 region) with an average length of 205bp from a sample of a chicken caecum. We note however, that our methodology is also applicable to longer 16S rRNA reads.

Before further analysis, we treated our dataset in the following way. To handle length variation among sequences, we trimmed our sequences to only be between the first and last conserved columns in the NAST [15] alignment to SILVA database [149]. We further removed any sequences less than 100bp long and any sequences that contained an unknown nucleotide (N). After cleanup our dataset had 21,646 sequences.

5.3.1 Multiple Alignment: Performance

We compared the effectiveness of the different alignment algorithms by using the likelihood values returned by maximum-likelihood phylogeny. In alignment of nucleotide sequences with secondary structure the aligners that are aware of the secondary structure generally outperform those that rely on sequence data alone [157], such as ClustalW [158] and MUSCLE [159]. In addition, these aligners scale poorly with dataset size [160]. Thus, we test the two commonly used 16S rRNA alignment algorithms: RDP [16]+Infernal and NAST in conjunction with the SILVA database [149]. Using ML phylogeny, we find log-likelihood scores of $-17,012$ and $-17,322$ for RDP+Infernal and NAST+SILVA, respectively, as obtained from the FastTree ML algorithm [143]. The merged alignment of RDP+Infernal with NAST+SILVA, described in the introduction, has a log-likelihood value of $-16,262$, representing an improvement over either of the two algorithms alone. When we perform further hand-curation of hypervariable regions of the 16S V3 in the merged alignment, we obtain a log-likelihood score of $-15,036$—reflecting the misalignments that can occur in the other automated procedure.

We can also take these different multiple alignments and cluster them in order to see how the OTU abundance results depend on the multiple alignment procedure. The OTU numbers after complete linkage clustering with radii 3%, 5% and 7% on seven different alignments are shown in Table 5.1. Here, "merge" refers to the merging of RDP+Infernal with NAST+SILVA. Note that running the aligners on sequences after quality processing produces thousands of OTUs. However, performing hand-trimming of sequence tails reduces the number of OTUs by an order of magnitude. This suggests that poorly curated alignments

Alignment method	Number of OTUs		
	3%	5%	7%
NAST+SILVA (on raw)	1141	646	406
RDP+Infernal (on raw)	3588	2313	1743
Merged (on raw)	3647	2297	1682
NAST+SILVA (on trimmed)	425	251	187
RDP+Infernal (on trimmed)	406	234	169
Merged (on trimmed)	393	227	165
Hand-curated	354	207	153

TABLE 5.1: Dependence of the number of OTUs on the alignment method used. The percentages indicate clustering radius. Trimmed sequences refer to sequences in which elementary hand-curation was performed (see introductory paragraphs of Results for more information). Merged refers to the multiple alignment that is a merging of the hypervariable regions aligned by NAST+SILVA regions with strong secondary structure conservation aligned by RDP+Infernal. See Introduction for more information. Note that crude hand-curation can reduce numbers of OTUs by a whole order of magnitude.

may overestimate microbial diversity.

5.3.2 Clustering: Performance

We compared the three clustering algorithms (complete linkage, FastGroup and ESPRIT) by running them on the hand curated alignment described in the previous section. We can visualize the effect of the choice of clustering algorithm by comparing rank abundance curves and cluster compositions. Rank abundance curves for the chicken caecum dataset are compared in Figure 5.4, for the 3% sequence difference clustering distance (and 1.5% FastGroup). As demonstrated by the curve, complete linkage clustering, ESPRIT and FastGroup 1.5% obtain the same shape of the curve, but FastGroup with 3% finds a very different one. This is because complete linkage at distance r corresponds to clusters where every element is at distance r to every other element in the cluster. On the other hand, FastGroup guarantees that every element is only at distance r from the chosen center of the cluster. This means that there may be elements in the same cluster that are at a distance of $2r$ from each other. Hence, r for FastGroup denotes the "radius" of the cluster, whereas r for complete linkage denotes the "diameter" of the cluster. Thus, FastGroup at 1.5% sequence distance can be compared to complete linkage and ESPRIT at 3%.

We find that FastGroup at 1.5% overestimates the number of OTUs in the sample. The binning in Figure 5.4 hides the fact that the number of OTUs found by FastGroup 1.5% is much larger than that of ESPRIT and complete linkage. FastGroup 1.5% finds 834 OTUs compared to complete linkage (354) and ESPRIT

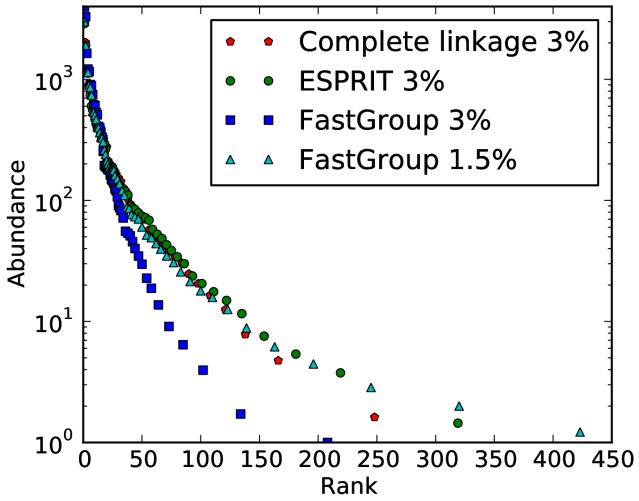


FIGURE 5.4: Rank abundance curves obtained with different algorithms and/or clustering distances. Notice that FastGroup with 1.5% sequence distance identifies a similar rank abundance curve to those of ESPRIT and complete linkage. However, it is not evident from the Figure that FastGroup identifies almost two times the number of OTUs than ESPRIT or complete linkage.

(434). Most of these extra OTUs are singletons. Of the 834 FastGroup OTUs, 440 are singleton OTUs. In comparison, complete linkage has 103 OTUs that are singletons out of total of 354. ESPRIT has 122 OTUs that are singletons out of 434 total. This is in accordance to the idea that is sketched in Figure 5.2, that a clustering algorithm such as FastGroup overestimates the number of OTUs.

We now evaluate the clustering quality via the CH index. For 3% clustering distance, complete linkage has a CH index of 167,771, whereas ESPRIT clustering has a CH index of 244. FastGroup with 1.5% clustering radius has CH index of 94,696. We note that complete linkage significantly outperforms other linkage clustering algorithms: nearest neighbor linkage (single linkage) got a CH index of 14,042 and average neighbor linkage got 23,512. We can also compare CH indices for clustering assignments that have roughly the same number of OTUs, rather than the same clustering distance. We find that complete linkage has CH indices between 140,000 and 160,000 for a range of clustering assignments with 200 to 300 OTUs. ESPRIT produced two clustering assignments in this range: first with 235 OTUs has a CH score of 280, and second with 303 OTUs has a CH score of 286. Finally, FastGroup (with 3% distance) got a CH score of 16,000 for a clustering assignment with 251 OTUs.

Another way to quantitatively judge the goodness of clustering is by comparing the OTU assignments to the structure of the maximum likelihood phylogenetic tree. To do this, we count the number of clades in a

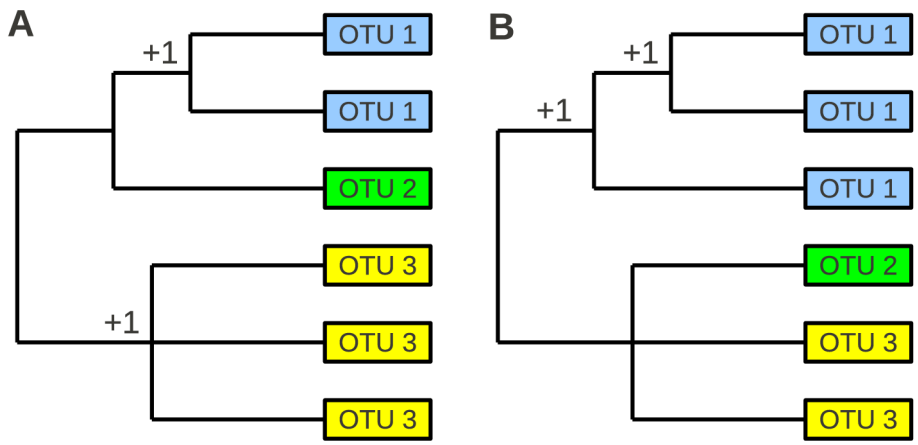


FIGURE 5.5: Sketch of the calculation of the number of clades with uniform OTUs. A phylogenetic tree with 2 different cluster (OTU) assignments is shown. The cluster assignment is indicated by OTU number and color. Both cluster assignments have 2 uniform clades (interior nodes indicated by +1). (a) The uniform clades are: one made up of two OTU 1 organisms, and one made up of three OTU 3 organisms. (b) The uniform clades are: one made up of two OTU 1 organisms and one made up of three OTU 1 organisms.

phylogenetic tree that contain only sequences of the same OTU (as determined by the clustering algorithm). We expect that a good clustering assignment will have many such clades. Two examples of this calculation are sketched out in Figure 5.5. We ran this calculation on 2 phylogenetic trees, one made by FastTree [143] (FT) and one made by RAxML [115] (RX), both inferred from our dataset described above. We find that complete linkage clustering has the most clades with uniform OTUs: 863 in FT and 698 in RX. Clustering with FastGroup (with 1.5% distance), we find 427 clades in FT and 367 in RX, whereas ESPRIT performs the most poorly: 6 clades in FT and 7 in RX.

We also explored if the rank abundance curves depend upon the clustering distance metric used. We find that the complete linkage clustering with the hand curated multiple alignment is very robust with respect to the choice of distance metric. In Figure 5.6a we compare the rank abundance curves (made by complete linkage) for three different distance metrics: Phylip DNADIST[147], percent sequence identity and distance along phylogenetic tree constructed by FastTree. We see that regardless of the choice of the distance metric, the shape of the rank abundance curve is conserved.

If we seek universal laws in the rank abundance data, we should expect that the shape of the rank abundance curve does not depend upon the particular clustering radius chosen. If instead rank abundance changes significantly with radius, that would imply that there is an interesting interplay between population dynamics and sequence distance. The complete linkage clustering with our hand curated multiple alignment

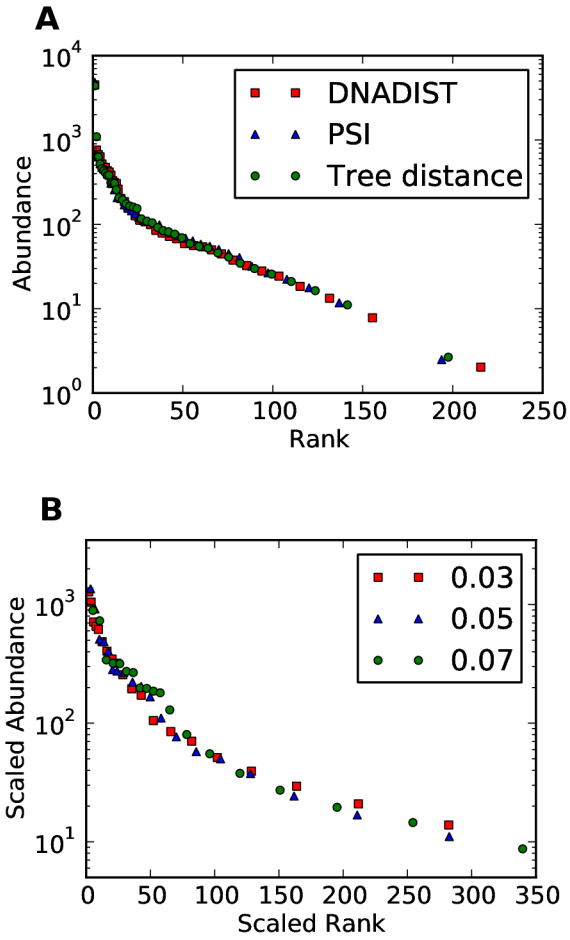


FIGURE 5.6: Two checks that should be used to verify quality of rank abundance curves. Both plots show rank abundance curves of the chicken caecum dataset. (a) Comparison of rank abundance curves for three clusterings using three different distance metrics. We compare the clusterings that produce 300 OTUs (which corresponds to different radii r for different metrics). (b) Rank abundance curve is robust if it does not change shape (functional form) when a different clustering radius is used. The rank abundance curves for different clustering radii all fall onto the same curve after rescaling the ranks to the same number of OTUs (while keeping area under the curve constant).

is found to be robust with respect to choice of clustering radius. As an example, see Figure 5.6b for the rank abundance curves of our chicken caecum microbial sample clustered at three different distances. By rescaling the axis of the rank abundance curves, while keeping areas under them constant, we can compare the functional forms (i.e. shapes) of the rank abundance curves. The figure shows that the chicken caecum microbial sample rank abundance seems to obey a universal law over a range of clustering distances.

5.4 Discussion of the Results

In the literature, the quality of data from pyrosequencing has been called into question [107, 161], especially with regard to its use in surveys of OTU diversity. Concern has been directed mostly at the experimental process of acquiring DNA sequences with high quality. Sogin *et al.* [95] showed that a number of heuristics can guarantee that per-base error rate of pyrosequencing is lower than that of Sanger sequencing while retaining more than 90% of data. Other artifacts that raised concern came from the shortness of pyrosequenced reads [103, 105]. Quince *et al.* [106] showed that reinterpreting pyrosequencing flowgrams via a maximum-likelihood scheme can lead to fewer OTUs. In this chapter we showed that a significant part of the discrepancy may arise from different computational analyses employed. Recent work [141] that has been similarly motivated has been commensurate with the conclusion that clustering is an important step in OTU analysis. In particular, they suggest that a preclustering step can help fix problems where deep sequencing overestimates species richness. Our work presents more general quantitative metrics that can be used as a standard for clustering programs. In addition, we find that calculating the log-likelihood of a maximum-likelihood phylogenetic tree is a good way to compare the quality of nucleotide alignments. Clustering quality index such as Colinski-Harabasz can even be used to verify what clustering radius is appropriate for a particular dataset. Our results that the multiple alignment and distance metrics can have a large effect on OTU abundances are also in agreement with recent work by Schloss [128].

In general, we found that multiple alignments can have a large influence on OTU abundance information, and the automated 16S rRNA alignment tools should be subjected to hand curation. Fast clustering tools such as ESPRIT do not make use of a multiple alignment and rely on k-mer heuristics to calculate pairwise distances between ungapped sequences. Our results show that such tools, intended to improve upon complete linkage, actually perform significantly worse. Hence, even with increasing dataset sizes, it is important to verify that the clustering method used performs no worse than complete linkage. We developed tools that ease the burden of performing hand curation and complete linkage of large contemporary datasets. These are available as supplementary software and are described in more detail in Sec. 5.5, Materials and Methods.

Finally, we summarize for the reader's convenience, step-by-step recommendations for handling a large 16S rRNA dataset, based on the analyses we have reported here. These are graphically illustrated in Figure

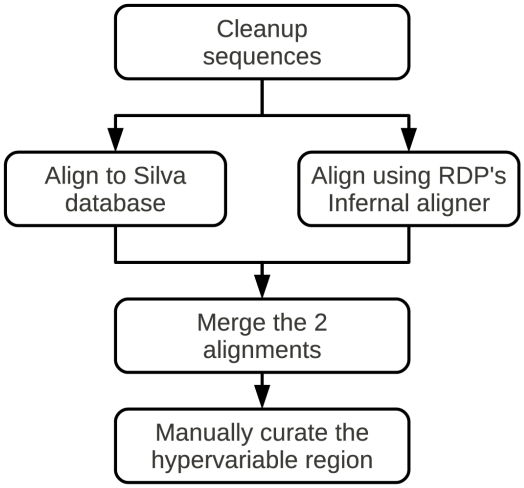


FIGURE 5.7: Diagram of our proposed 16S rRNA alignment pipeline, TORNADO. After the preliminary clean up step, we align the sequences in two different ways. First, we use Mothur [139] to align our sequences to the SILVA [149] database. Second, we align using Ribosomal Database Project's front end [16] to the Infernal aligner [17]. We then merge the two, using Infernal's secondary-structure-aware alignments and SILVA's alignment of hypervariable region. Finally, we manually curate the hypervariable regions, using a helper tool, splicer, we developed (see Figure 5.8).

5.7.

1. *Quality Processing:* Remove short reads and sequences with unknown nucleotides (N). Make an alignment to the SILVA database [149], via NAST [15] as implemented by Mothur [139]. Trim sequences to be between the first and last strongly conserved columns in this alignment.

2. *Alignment:* From the trimmed dataset, produce another alignment through RDP pipeline's [16] front end to the Infernal aligner [17]. Merge the two alignments (NAST+SILVA with RDP+Infernal) using the tool that's a part of the TORNADO pipeline at `http://tornado.igb.uiuc.edu`. Further hand-optimize hypervariable regions of the reads by using the tool available on the website above.

3. *Cluster:* Cluster the dataset using the complete linkage tool available on the website above.

Further analysis can be performed by calculating estimators in Mothur [139], or by estimating phylogenetic trees via RAxML [115] or FastTree [143].

5.5 Materials and Methods

5.5.1 V3 rRNA amplicon sequencing.

We used the V3 rRNA sequences from the chicken caecum from batch B of a previous study [162]. PCR specific primers flanking the V3 hypervariable region of bacterial 16S rRNA were used to generate PCR products for pyrosequence analysis. The forward fusion primers for pyrosequencing included 454 Life Sciences A adapter, and barcode *A* fused to the 5′ end of the V3 primer 341F (5′ gcctccctcgcgccatcag-ACGAGTGCGT-CCTACGGAGGCAGCAG3′) or with barcode *B* (5′ gcctccctcgcgccatcag-ACGCTCGACA-CCTACGGAGGCAGCAG3′). The reverse fusion primer included 454 Life Sciences B adapter fused to 5′ end of V3 primer 534R (5′ gccttgccagcccgctcag- ATTACCGCGGCTGCTGG3′). Cycling conditions (20 cycles) were; initial denaturation at 94 Ph.D.C for 5 min; 20 cycles of 94 Ph.D.C 30 s, 60 Ph.D.C 30 s and 72 Ph.D.C 30 s; then 72 Ph.D.C 7 min for final extension. The amplicon products were cleaned using PCR purification clean-up kit and SPRI size exclusion beads. The quality of products was assessed using a Bioanalyzer using DNA1000 chip. The fragments in amplicon libraries were subjected to a single pyrosequence run using a 454 Life Science Genome Sequencer GS FLX (Roy J. Carver Biotechnology Center, University of Illinois). The resulting dataset had 22953 sequences of average length 204.7bp. Before further analysis was performed, we performed basic filtering. We removed all sequences that were shorter than 100bp reducing the number of sequences to 21646. The sequences have been uploaded to GenBank (accession numbers HQ293272-HQ315544).

5.5.2 Multiple Alignments

We compared 4 different alignment methods as illustrated in Figure 5.1. (1) We fed the sequences into Infernal [17] with bacterial secondary structure template as provided by RDP [16]. (2) We aligned the sequences to the SILVA database [149] using the NAST[15] algorithm as implemented by Mothur [139, 163] (align.seqs command). (3) The results of (1) and (2) were then fed into a merger script we have made available on the Web at `http://tornado.igb.uiuc.edu`. (4) The merged data sets' hypervariable regions were then hand curated using `splicer`, a tool we developed and made available on the Web as part of our pipeline TORNADO at `http://tornado.igb.uiuc.edu`. This tool allowed us to greatly reduce the number of unique snippets of the hypervariable region of V3 down from 21,646 to about 200, by

cutting the longest hypervariable subregion from the alignment, and then dereplicating it. These snippets of sequences in the hypervariable subregion ranged from 1bp to about 30bp. This meant that we only needed to hand-curate 200 short snippets to handle the alignment of the hypervariable region. These snippets were separated into two groups according to their secondary structure: loop, and stem-loop-stem. We used RNAfold web server [164–167] to verify the structure. The two groups were then hand curated and merged back into the complete multiple alignment using the `splicer merge` command. For clarity, the process of using `splicer` is described in Figure 5.8. All multiple alignments are available at `http://tornado.igb.uiuc.edu`.

5.5.3 Likelihood Scores

Each data set described in the previous section was dereplicated producing 2215 clones each. Likelihood scores were then computed for each dataset using FastTree 2.1.1 with command line parameters `-gamma -nt -gtr`.

5.5.4 Distance metrics

We compared 3 different distance metrics to generate Figure 5.6a. (1) Phylip DNADIST 3.67 [147] with default model parameters. (2) Percent sequence difference calculated using a program we developed, `psi-distance`, available at `http://tornado.igb.uiuc.edu`. The program constructs pairwise differences by calculating the number of letters that are different between every two sequence (gap is considered a letter). This number is then divided by the average of the ungapped lengths of the two sequences compared [145]. (3) Tree distance calculated from the phylogenetic tree calculated by FastTree in the previous section. The tree distances were acquired by calculating tree branch lengths from the Newick formatted tree using the `tree-distance` program we developed, available at `http://tornado.igb.uiuc.edu`.

5.5.5 Clustering algorithm

Three different clustering algorithms were compared, all on the hand-curated dataset. (1) Complete linkage clustering with furthest neighbors, as implemented in `c-linkage`, a program we developed that is available at `http://tornado.igb.uiuc.edu`. We tested that the program produces the same results as Mothur, but much faster and with less memory usage since it works in $O(N^2 \log N)$. For a comparison of running times

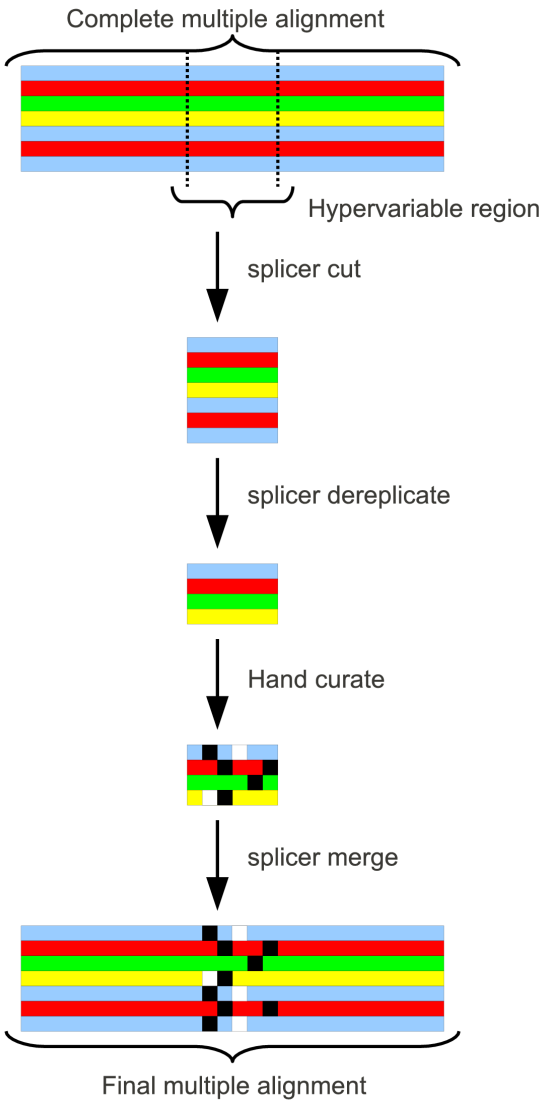


FIGURE 5.8: Using splicer, a part of the TORNADO pipeline, to perform hand curation. Dereplicating the hypervariable region significantly reduces the effective number of snippets of sequences one needs to hand curate (4 instead of 6 in this example). In our dataset of around 20,000 sequences, there were only around 200 unique sequence snippets in the hypervariable region varying in length between 1 and 30 bp.

of `c-linkage` and Mothur version 1.12.3, when clustering up to a clustering cutoff of 10% see Figure 5.9. (2) FastGroup[153] with no trimming, PSI difference of 97% with gaps. (3) ESPRIT[154], for which the dataset was first degapped.

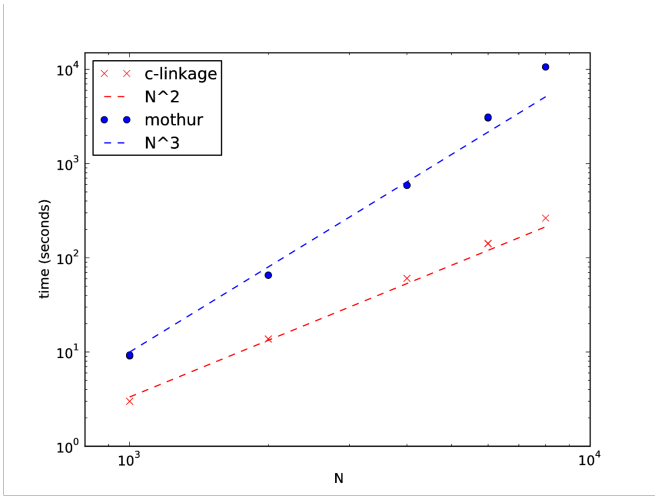


FIGURE 5.9: Comparison of running times of c-linkage with the running times of Mothur. The two programs were benchmarked on artificial datasets of 1000, 2000, 4000, 6000 and 8000 elements. The scripts used to generate these datasets and run the benchmarks are available at `http://tornado.igb.uiuc.edu`.

5.5.6 Cluster Metric

We evaluated the quality of the clustering by calculating the Calinski-Harabasz index[142, 156, 168]. The implementation of the program that calculates the index is available at `http://tornado.igb.uiuc.edu`.

Chapter 6

Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes

The theoretical description of the forces that shape ecological communities focus around two classes of models. In niche theory, deterministic interactions between species, individuals and the environment are considered the dominant factor, whereas in neutral theory, stochastic forces, such as demographic noise, speciation and immigration are dominant. Species abundance distributions predicted by the two classes of theory are difficult to distinguish empirically, making it problematic to deduce ecological dynamics from typical measures of diversity and community structure. Here we show that the fusion of species abundance data with genome-derived measures of evolutionary distance can provide a clear indication of ecological dynamics, capable of quantifying the relative roles played by niche and neutral forces. We apply this technique to six gastrointestinal microbiomes drawn from three different domesticated vertebrates, using high resolution surveys of microbial species abundance obtained from carefully curated deep 16S rRNA hypervariable tag sequencing data. Although the species abundance patterns are seemingly well fit by the neutral theory of metacommunity assembly, we show that this theory cannot account for the evolutionary patterns in the genomic data; moreover our analyses strongly suggest that these microbiomes have in fact been assembled through processes that involve a significant non-neutral (niche) contribution. Our results demonstrate that high-resolution genomics can remove the ambiguities of process inference inherent in classical ecological

measures, and permits quantification of the forces shaping complex microbial communities.

6.1 Introduction

Ecological species distributions are determined by the interplay between environmental factors and evolutionary processes. In classical ecological theory, niches, characterized, for example, by nutrients and other environmental factors, determine species abundance distributions and populations primarily through deterministic partitioning of resources amongst species [169]. Species populations are limited by niche carrying capacity, rather than interspecies competition, thus tending to promote coexistence [170]. In niche theory, diversity is determined primarily by the number of available niches, raising the issue of how to account quantitatively for the apparent observed diversity [171–174] from well-documented instances of niche differences [175].

An alternative perspective is the class of neutral theories, in which species are functionally equivalent, and stochastic factors such as immigration, birth-death processes and speciation are the primary drivers of ecological diversity and community structure [176–181]. This class of models has been reported to be capable of accurate predictions for the species abundance distributions in (e.g.) riverine fish populations [182] or microbial populations [183], in addition to the early successes in forest ecosystems, a planktonic copepod community, and a bat community in Barro Colorado Island (BCI) [178]. However, the methodology used in such comparisons is contentious when examined carefully [184, 185], with sampling issues, parameter estimation, and model definition being some of the key factors that require careful attention. The assumptions of neutral theory, in particular functional equivalence, are not transparently biological [186], and additionally have been criticized on a variety of empirical grounds [187, 188], including the predictions for species lifetimes, speciation rates and the incidence of rare species [189]. Other technical assumptions, for example that the number of individuals competing for a resource is a constant (the "zero-sum" assumption), may be unrealistic, but can be extended or relaxed [181, 190, 191]. Perhaps a more useful insight into the applicability of neutral theory comes from considering the interplay between niche stabilization mechanisms and fitness [192]. A recent study of a sagebrush steppe community, where strong niche stabilization mechanisms were identified even in the presence of apparently small fitness differences [193], underscores the fact that weak functional inequivalence need not necessarily mean that niche dynamics are negligible. On the other hand, a study that attempted to infer pairwise interaction strengths among the most abundant species in the

BCI site found that interspecies interactions were much weaker than intraspecies one, in apparent agreement with neutral assumptions [194].

Despite their fundamental differences, and the plethora of studies nominally supporting each side of the niche-neutral dichotomy, these theories predict species abundance distributions that are difficult to distinguish empirically [173, 195], with similar mathematical properties for asymptotically large diversity [196]. The inverse problem of inferring ecological dynamics from measures of diversity does not appear to have a unique solution, either theoretically or empirically. Accordingly, a more nuanced perspective has arisen [170, 187, 197], in which elements of both types of theory may contribute to a proper description of the ecological dynamics, and a variety of mathematical frameworks for accomplishing this type of synthesis have recently appeared [194, 198–203]. Nevertheless, it remains an open question as to how to properly characterize community dynamics, and how to usefully quantify the relative roles of niche and neutral processes in the evolutionary dynamics of ecosystems.

These questions are of particular relevance to microbial communities, which play functionally important roles in ecosystems, but are typically rich in diversity, suggesting the presence of sub-populations shaped primarily by stochastic forces. Such communities would not be expected to represent endmembers of the niche-neutral continuum, and quantification of their structuring process represents a complex problem that has recently attracted attention. Most studies find evidence for a mixture of neutral and niche processes in microbial community assembly [204–208]. These seem to arise for different physical reasons. One indication is that the neutrally-assembling taxa are generalist microbes, that can exist in a wide variety of environments [206], whereas the niche portion of the microbiota are adapted to the media conditions [209]. There are also indications that that microorganismal coocurence patterns are shaped by the same processes and interactions that shape macroorganismal coocurence patterns [210].

In this chapter we propose a methodology for addressing the problem of quantifying the relative role of niche and neutral processes in structuring microbial communities, by fusing measures of abundance with phylogenetic information. The merging of classical ecological measures with phylogenetic analysis is growing in importance, but is still in its infancy [211–215]. The method presented here is particularly applicable to uncultured microbial communities that are characterized by a high level of diversity, and are amenable to modern metagenomic tools, such as pyrosequencing.

In order to explain the basic idea of how we quantify an ecosystem on the niche-neutral continuum, it

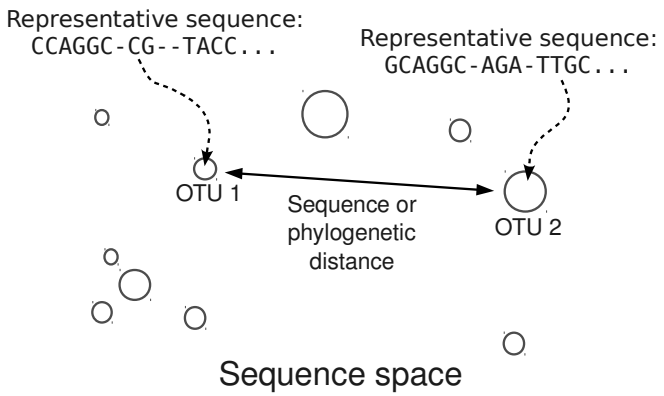


FIGURE 6.1: Sketch of the starting point for a metagenomic analysis of an environment. Circles indicate OTUs, and abundance (number of sequences within the OTU) is labeled by the size of the circle. A representative sequence is associated with each OTU. The OTUs are embedded in a sequence space such that the distance between the circles in the sequence space corresponds to e.g. sequence or phylogenetic distance between the representatives.

is necessary to recall how microbiomes can be probed by genomic methods. The first step in an ecological study of a microbiome, following sequencing, cleanup and alignment, is the assignment of sampled sequences into Operational Taxonomic Units (OTUs) through a clustering process [216]. The OTUs are then used as a proxy for estimating microbial species abundance [104]. The OTU data are two-fold. On the one hand, the OTUs have relative abundances that are estimations of the species' abundances in the environment. On the other hand, OTUs also have representative sequences associated with them. Typically a representative sequence of an OTU is the most abundant of the identical clones within the OTU, and also it is more than 97% similar to every other sequence within that OTU. This genomic data associated with the representative sequence allows us to think of OTUs as points in a sequence space as illustrated in Fig. 6.1. We can think of distances between points in this space as corresponding to the phylogenetic or sequence distances between the sequences in these OTUs.

This cloud of points in high-dimensional sequence space can also be labeled by OTU abundance. In our work, this is determined by sequence abundance (after every effort has been made to account for artifacts), but in principle OTU abundance labels could be obtained from any other source, such as Q-PCR. In this space, we can categorize the OTUs into two sorts: the most abundant OTUs (which we term "modal" OTUs, and define this precisely below) and the other, less abundant, OTUs (which we term "rare" OTUs, and define this precisely below). The correlations between the modal and rare OTUs will depend upon the evolutionary dynamics, and in fact exhibit sharp mathematical differences that can be used to discriminate different putative dynamics. To see the essential idea, we will now explain how this would work in two caricatures

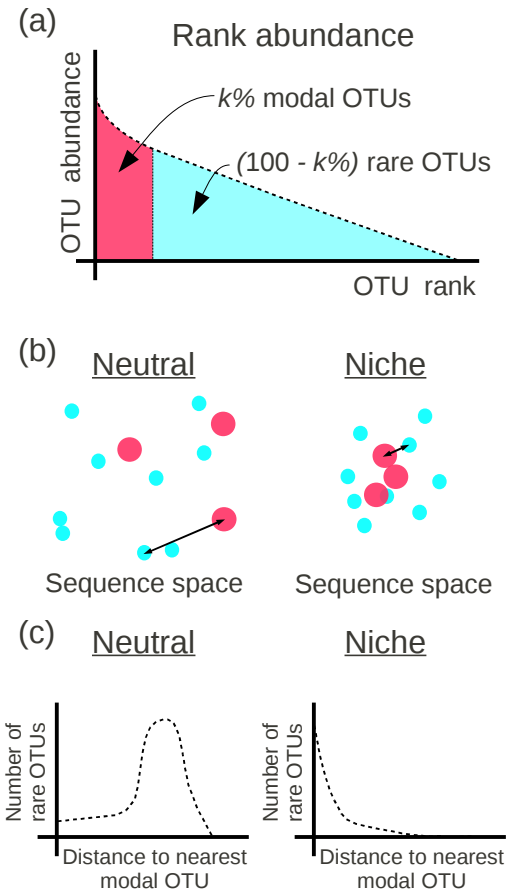


FIGURE 6.2: (a) Classification of the OTUs into two groups based on the rank abundance. The top $k\%$ of OTUs are labeled modal, whereas the remainder of the OTUs are labeled rare. (b) Sketch of the neutral and niche evolution processes in sequence space. Light blue OTUs are rare, whereas red OTUs are modal. For the neutral process, the average distance of a rare OTU to its closest modal OTU is large (indicated by the arrow). For the niche process, this distance is much smaller since rare OTUs cluster about the modal OTUs which define the niches. (c) Sketch of the expected distributions of distance to the closest modal OTU. For the neutral process, this distribution is peaked around some non-zero distance, which is close to the average distance between the OTUs in the dataset. In the niche process, the distribution monotonically decays with distance since the rare OTUs are attracted to the niches.

of ecosystem dynamics: a simplified neutral model and a simplified niche model. A significantly more elaborate analysis is carried out below, in the main body of this chapter, but the key concepts are captured by these simplified models.

First, suppose that the evolutionary dynamics is itself neutral, so that the rare and modal OTUs are distributed at random in the high-dimensional sequence space. We are going to be interested in measuring the distances between sequences corresponding to different OTUs, and comparing their similarity. Let us assume that the sequences being analyzed are all of the same length, containing L nucleotide bases from the usual 4-letter alphabet (ACGT); here we are ignoring real life complications such as insertions, deletions

and gaps. We label the sequences by S_α^i, where $\alpha = 1 \ldots L$ labels position along the sequence and i labels the OTU; S_α^i can take the values 1,2,3,4 corresponding to the alphabet of bases ACGT. We define the normalized Hamming distance H_{ij} between two sequences i and j as the fraction of bases in i that are different from the base in the corresponding position in j:

$$H_{ij} \equiv \frac{1}{L} \sum_{\alpha=1}^{L} (1 - \delta(S_\alpha^i - S_\alpha^j)) \tag{6.1}$$

where δ denotes the Kronecker delta. The mean $\langle H \rangle$ of H_{ij} averaged over a large sample of random sequences would be $3/4$, because there is a $1/4$ chance that two bases at the same position are identical. Thus, the probability distribution of H would be expected to be a roughly bell-shaped curve, peaked around $H = 3/4$, with a width dependent on the number of sequences. In practice, there are complications due to insertions, deletions and gaps, but most importantly, conserved positions. Bases that are highly conserved cannot be appropriately modeled as being chosen randomly from the alphabet. This can be taken into account by simply restricting the above analysis to bases that are strongly non-conserved: let us call the number of highly conserved bases $M < L$, so that the expected value of H will now be reduced by the fraction of conserved bases: $\langle H \rangle = 3(L - M)/4L$. Thus, taking into account conservation, the bell-shaped curve will shift its peak to a smaller value of H. In the data presented below, we found that $L \sim 200$ and $M \sim 160$, so that the distribution of H should be peaked at about 0.15, in the case of a neutral system. Now consider a subset $\{E_k\}$ of distances $\left\{H_{ij}\right\}$. For each "rare" OTU k, we rank all of the distances between OTU k and each "modal" OTU l. Then, we select the shortest such distance and label it E_k. In this way, the set $\{E_k\}$ is the set of distances of "rare" OTUs to their nearest niche neighbor. For the above case where the evolutionary dynamics is neutral-like, the distribution of E is also a bell-shaped curve like the distribution of H. However, its mean is slightly shifted to the smaller values, and its standard deviation is smaller (because $\{E\}$ is the subset of shortest distances from the set of $\{H\}$). In other words, $\langle E \rangle < \langle H \rangle$.

Second, let us consider a caricature of a system that is dominated by niche dynamics. In the extreme (and unrealistic) case where there is only one niche, occupied by one particular modal OTU, the probability distribution of E will be a delta distribution peaked at $E = 0$. In a more realistic model, where there is a cloud of rare OTUs surrounding the modal OTU, having evolved from it by a few point mutations, one would expect the probability distribution of E to be peaked at $E = 0$, and then to monotonically decrease for $E > 0$. In the case of a system with several niches, the probability distribution for E will

be somewhat more complicated, because one needs to calculate the normalized Hamming distance from each rare OTU to the nearest modal OTU, and this requires making a Voronoi polyhedron construction in sequence space. Nevertheless, for small values of E, the probability distribution will be dominated by the single niche argument given above, and the functional form will be unchanged: peaked at the origin and monotonically decreasing for $E > 0$. These two caricatures for simplified models of ecosystem structure are sketched in Fig. 6.2, and show that there are clear and distinct signatures arising from the nature of the processes that have structured the community.

In the remainder of this chapter, we numerically evaluate the metric for model systems in order to quantitatively and concretely confirm the above heuristic description. We then describe how we have implemented these ideas in a proof-of-principle study of vertebrate gastrointestinal microbiomes. These experimental systems were chosen, not only because of the growing recognition of the importance of microbiomes as a determinant of host health[217], but also because these are systems that have high diversity, and are likely to be shaped both by stochastic and niche processes. Indeed, as we will see, they can be well-described naively by neutral theory, although in fact niche processes play a fundamental role in structuring these communities.

6.2 Model calculations

In this section we evaluate our metric on model systems parametrized by a single parameter, α, the proportion of the system undergoing a niche dynamic. We perform 5000 Monte Carlo simulations of the following process. We simulate N OTUs (here $N = 1000$) each with representative sequences of length $L = 200$. A subset αN ($0 \leq \alpha \leq 1$) of the OTUs undergo a niche dynamic in the following way. A single random OTU is chosen to be the center of the niche. The remainder of the $\alpha N - 1$ OTUs (niche OTUs) are are generated by performing random mutations of the genome of the OTU representing the niche center. The placement and number of the mutations were chosen randomly in the following way. Placements of mutations were sampled uniformly (without replacement) across the entire genome. The number of mutations for each of the niche OTUs was sampled from an exponential distribution thereby modeling the evolution of OTUs under multiplicative fitness pressure (larger number of mutations corresponds to smaller fitness, and hence smaller abundance of OTU). The remaining $(1 - \alpha)N$ OTUs (neutral OTUs) are randomly distributed throughout the sequence space, and they represent the sequences undergoing dynamics under no evolutionary pressure (neutral dynamics).

Each OTU in the model system is associated with an abundance. The abundances of neutral OTUs are randomly sampled from an exponential distribution. (In the Hubbell Neutral Model, the OTU rank abundances are exponentially distributed.) On the other hand, the abundance of niche OTUs exponentially scales with their closeness to the niche:

$$N_i = A \exp(-d_i) \tag{6.2}$$

where N_i is the abundance of OTU i and d_i is the distance from the OTU to the center of the niche (in sequence space). The results of our metric, the distributions of $\{E_k\}$ are shown in Fig. 6.4 for 3 model systems characterized by values of $\alpha = 0, 0.5$ and 1. We see that the heuristic arguments we described in the previous section and sketched out in Fig. 6.1(c) are consistent with these model numerical calculations.

It is instructive to demonstrate the effects of two factors on our metric, in order to highlight some of the mathematical considerations that went into the design of the metric, in particular our use of an extremal measure (the shortest distance aspect of our metric) and the influence of sampled abundance distributions. First we demonstrate the role of extremality introduced by choosing the subset $\{E\}$. Instead, if we choose to plot the distribution of $\{H\}$ we obtain qualitatively the same results for neutral-like models (compare models 1 and 2 in Fig. 6.3). However, for niche-like models, the peak at zero moves to a nonzero peak which corresponds to the average size of the niche (compare models 5 and 6 in Fig. 6.3). Thus, the choice of an extremal measure is important in making sure that the endmember distributions (pure niche, pure neutral) are clearly distinct.

Second, we demonstrate what might appear at first to be a rather counter-intuitive fact: the distribution of distances is only weakly dependent on the abundance distribution of the OTUs. If the abundance of an OTU k is N_k then we could imagine modifying our procedure by weighting the contribution of E_k in the distribution $\{E\}$ by a factor of N_k. Such a weighting introduces no change whatsoever to the neutral dataset (compare models 2 and 4 in Fig. 6.3), and no qualitative change in the niche dataset (models 6 and 8 in Fig. 6.3). Finally, we can also weigh the distribution of $\{H\}$ in such a way that each distance H_{ij} between OTUs i and j gets weighted by a factor of $N_i N_j$. The results are exactly the same as with no weighing for the neutral dataset (compare models 1 and 3 in Fig. 6.3) and qualitatively the same for the niche dataset (compare models 5 and 7 in Fig. 6.3).

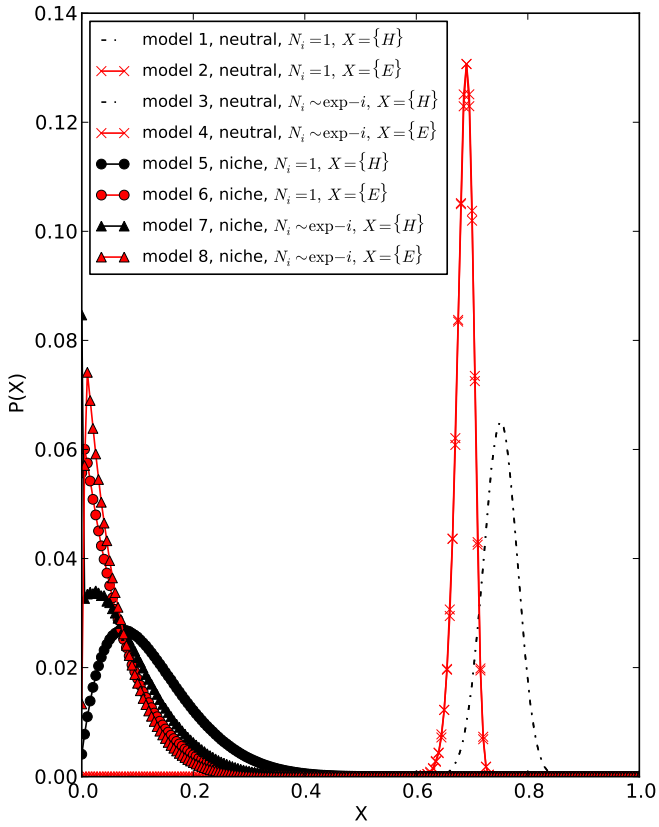


FIGURE 6.3: Explicit numerical calculations of our metric on 8 model systems. In these systems, we study the difference between the effects of the metric on neutral (models 1-4) and niche model systems (models 5-8). We also study the effect of choosing the closest distance (even-numbered models) compared to considering all distances (odd-numbered models). Finally, we consider the weighted models (3-4 and 7-8) versus the unweighted ones (1-2 and 5-6).

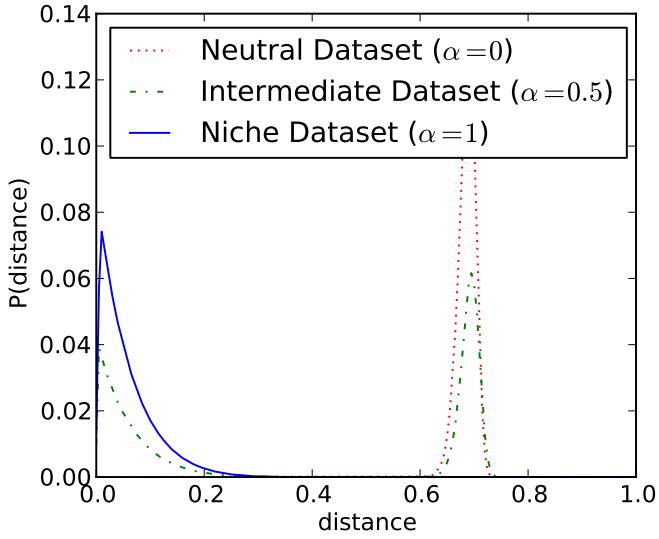


FIGURE 6.4: The results of our metric, the distributions of E shown for a fully Niche-like model dataset ($\alpha = 1$), a fully Neutral-like model dataset ($\alpha = 0$) and an intermediate dataset ($\alpha = 0.5$). The results shown are the average of 5000 Monte Carlo simulations for each dataset.

	# reads	Unique reads	Average length	Aligned width	# OTUs at 3%
Swine feces 1	33283	14122	165.0	420	1509
Swine feces 2	36254	16198	175.3	418	1856
Cattle rumen 1	31201	18264	180.7	471	2580
Cattle rumen 2	19642	10074	183.6	385	1509
Chicken caecum 1	17585	2151	136.5	310	396
Chicken caecum 94	21646	2223	138.9	332	354

TABLE 6.1: Summary statistics of our six datasets.

6.3 Results

We performed a pyrosequencing study of the gastrointestinal (GI) microbiomes of 3 pairs of domesticated vertebrates: 2 swine, 2 cattle and 2 chickens. These pairs of organisms were chosen as pilots for probing specific microbiome issues of relevance to animal science. In particular, we attempted a comparative study looking at the effects of diet on identically cloned swine, and the effects of a microbial challenge on two identically-raised chickens. For the purposes of this chapter, these comparisons and the outcomes of the experiments are not of interest: full details of the comparisons and other studies will be published elsewhere. In this study, two genetically identical cloned swine were fed different diets and then their fecal samples were collected for sequencing. Cattle rumen 1 and cattle rumen 2 were rumen fistula sampled at 0 and 8 hours after feeding, respectively [218]. Chicken caecum 94 was inoculated with *Campylobacter jejuni* one week prior to caecal sampling. Chicken caecum 1 was kept under the same conditions but without oral gavage of *C. jejuni* [162]. See the Methods for details regarding the laboratory protocols. The GI Samples were subjected to deep hypervariable 16S rRNA tag sequencing using a 454 Life Science Genome Sequencer GS FLX [104]. Table 6.1 shows the average read length and number of reads obtained for each sample.

Following their acquisition, we aligned the pyrosequenced reads using NAST [15] to the SILVA [149] database. We also aligned the reads using RDP's frontend [16] to the Infernal [17] structural aligner. For each dataset, the NAST+SILVA and RDP+Infernal multiple alignments were merged and hand curated using the methodology and tools described in Sipos *et al.* [216]. Short reads and sequences with unknown nucleotides were removed. Spurious "tails" in the multiple alignment, sequences that extend beyond the

	Simpson diversity	Shannon diversity	Jackknife richness	ACE richness	Chao1 richness
Swine	0.0070	5.8	2000	1472	1540
feces 1	±0.0003	±0.02	±260	±55	±150
Swine	0.0068	5.9	2300	1633	1720
feces 2	±0.0003	±0.02	±300	±53	±150
Cattle	0.0044	6.3	3300	3070	2640
rumen 1	±0.0002	±0.02	±260	±88	±190
Cattle	0.0110	5.9	2070	1818	1830
rumen 2	±0.0006	±0.03	±110	±62	±130
Chicken	0.084	4	770	655	620
caecum 1	±0.003	±0.03	±120	±75	±150
Chicken	0.046	3.9	560	426	460
caecum 94	±0.001	±0.02	±90	±57	±100

TABLE 6.2: Summary diversity metrics of our six datasets.

region of 16S common to all the sequences in the dataset, were also removed. Distance matrices were generated from the multiple alignments, and were then fed to a complete linkage clustering algorithm to generate the OTUs. The careful multiple alignment procedure led to a vast reduction in the number of resulting OTUs in the datasets as previously reported in Sipos *et al.* [216]. See Table 6.2 for species diversity and richness metrics for each of the 6 GI microbiome samples. Rarefaction curves show how the number of sampled OTUs varies as a function of the number of organisms sampled. Our rarefaction curves are shown in Fig. 6.5 for each of the 6 datasets.

We plotted the abundances of the OTUs for each of the 6 datasets in our study, and we find a very good agreement with the Neutral Model. These are displayed in rank-abundance form in Fig. 6.9, and in alternative forms in Figs. 6.6 and 6.7. The early ranks (high abundance OTUs) show some systematic deviation from the abundances expected from neutral theory but at face value, these results are consistent with the majority portion (thousands) of the OTUs evolving in the absence of any apparent selection acting on the individual OTUs. Given all the factors that influence the gastrointestinal microbiome [98, 219–223], and the reproducible, thereby seemingly host-selected, microbial abundances [224], it seems counterintuitive that there should be no apparent selection for the vast majority of OTUs in the exponential tail of the rank abundance. However, if we compare taxonomic assignments of microbes across each pair of animals in our study (Fig. 6.8), we find that there is a correlation between the relative abundances of taxa in members of each animal pair. Namely, we observe that the most abundant taxonomic orders are the same for each animal pair (Clostridiales for swine and chickens, and Pseudomonadales for cattle). This correlation also extends to other taxonomic orders. Hence, our dataset indicates that certain taxa are favored more than others within

the GI tract of these 6 vertebrates.

We now attempt to resolve this apparent contradiction, namely that the Neutral Theory fits the rank abundance patterns well, with only 2 fitting parameters, even though the taxonomic data suggests Niche selection. In order to do this, we must turn our attention to other information contained within the pyrosequenced reads. As shown in Fig. 6.1 the OTUs with their characteristic sequences and associated abundances form patterns within a high-dimensional space. Each read constitutes a point in this space, defined by its nucleotide sequence. One way in which we can attempt to comprehend the structure of this space is through dimensional reduction. We use Principal Component Analysis (PCA) in order to place the OTUs into a 2 dimensional space spanned by the two principal components. We perform a weighted version of PCA [225] where we assign a weight to the OTUs proportional to their abundance. The resulting patterns in the space of two principal components are shown in Fig. 6.10. Each circle in the figure is an OTU and the circles' size and color indicates the logarithm of the OTU abundance.

As a control, we generate datasets of artificially generated sequences (hereafter referred to as neutral datasets). We generate a neutral dataset for each of the 6 experimental datasets to facilitate a 1-to-1 comparison. Each neutral dataset is constructed in a way such that it has the same number of OTUs and the same OTU abundance distribution as the associated experimental dataset. However, the representative sequence for each OTU is artificially generated and has a randomized sequence, with constraint such that it has the same sequence statistics as the original dataset (probability of observing a nucleotide at a position in the multiple alignment) and column conservation. This ensures that the sequences are randomly distributed along a realistic sub-manifold of sequence space (the subset of 16S sequences that are allowed by secondary structure). We then run the PCA on the neutral datasets (Fig. 6.11). Comparing Figs. 6.10 and 6.11, we notice the following pattern in the experimental GI data: the low-abundance OTUs cluster around the high-abundance OTUs in the dimensionally reduced space. In the neutral datasets, this is not observed, instead the PCA distributes the OTUs approximately uniformly in the dimensionally-reduced space.

We now formulate a heuristic to clearly discriminate between the randomly assembled model sequences and those assembled from a niche-driven process. On a rank-abundance curve, we label the $k\%$ of the most abundant OTUs as modal OTUs. We label the remaining OTUs as rare OTUs (Fig. 6.2(a)). Instead of using the whole-dataset rank-abundance curve, one can also use per-order rank-abundance curves if additional resolution is necessary. Once modal and rare OTUs have been assigned, for each rare OTU we compute the

distance to the modal OTU that is closest to it. The motivation behind this heuristic is the following. The spread pattern of sequence abundances gives us an indication of whether organisms are evolving neutrally or toward defined niches. In long time behavior, neutral evolution leads to the expectation that organisms have an equal chance of being anywhere in this space. Niche selection, however, suggests a very biased distribution of organisms. In particular, organisms would be densely clustered about the local optimum for each niche (Fig. 6.2(b)). These two scenarios lead to very different distributions of distance to nearest niche. If the OTUs are undergoing a niche-driven dynamic, then the rare OTUs will tend to drop off exponentially in abundance around the modal OTUs. If on the other hand, the OTUs have been sampled from a community shaped by neutral evolutionary dynamics, then the rare OTUs' distance to closest modal OTU will be peaked around some non-zero distance that is the average distance between any two OTUs in the dataset (Fig. 6.2(c)).

We apply the above analysis to the case of gastrointestinal microbiome datasets of the 6 vertebrates. The results are summarized in Fig. 6.14. In this figure, the blue bars indicate the results of our metric applied to experimental data. The dashed red lines indicate the results of the metric applied to a dataset of sequences that were randomized in the way described above. The results indicate that the GI tracts of the 6 vertebrates largely undergo niche dynamics, with the possible exception of a subpopulation of the chicken GI tracts. The chicken datasets have a small non-zero peak corresponding to the average distance between sequences chosen at random. Our study indicates that the sequences within this peak may be undergoing neutral dynamics. The results that we obtain are robust in that they do not qualitatively depend upon the choice of the cutoff k. In Fig. 6.13 we show the metric for $k = 5\%$ and $k = 7\%$. Similarly, the results of the metric on model systems are virtually unchanged when k is changed between 2% and 10% (Fig. 6.12) indicating robustness. Whereas our metric is robust in this way, the reader is reminded that phylogenetic resolution is nevertheless important: some niches may appear as a single OTUs at 97% sequence identity.

6.4 Discussion

In this work, we set out to construct genomic-based measures of ecosystem diversity and abundance that can provide evidence for process. We focused on understanding the processes that structure microbial communities, because these play functionally important roles in many ecosystems, yet are rich in diversity. Thus, such systems would *a priori* be expected to contain at least sub-populations shaped primarily by stochastic

forces. The dual features of high diversity and foundational role functionally in their host ecosystem suggests that microbial communities would not be simple to characterize as either niche or neutral. At the same time theoretical arguments suggest that such high-diversity communities might appear, for fundamental statistical reasons, as neutral.

We succeeded in creating a quantitative metric that fuses abundance and genomic data in order to determine whether an ecological system is dominated by neutral evolution or by niche selection. The key concept was to explore the correlations and associated probability distributions between the most abundant members of the community and the long, low abundance tail members. We showed that the signature of the probability distribution describing the distance in genomic sequence space from each rare OTU to the nearest modal OTU provided a signature of the strength of niche dynamics. We tested this construct on large datasets from 6 animal gastrointestinal tract microbiomes, finding in all cases that the results are inconsistent with neutral assembly. We conclude that niche selection largely dominates within the GI microbiome, despite the fact that the rank abundance patterns are apparently well-modeled by Neutral Theory.

Our results provide firm evidence from an empirical dataset that apparently neutral patterns of diversity and abundance can arise from niche-dominated dynamics, in agreement with earlier theoretical expectations [170, 173, 187, 195–197]. Our results establish definitively that simple ecological measures need to be, and can be, augmented by genomic data in order to provide insight into the processes that structure communities.

6.5 Materials and Methods

6.5.1 Sample Preparation

All procedures involving animals were approved by the Institutional Animal Care and Use Committee of the University of Illinois. For each animal, we used two different samples for our test that vary in some aspect such as diet or sampling times. The Duroc sow (2-14; TJ Tabasco) was used as the genomic template for producing cloned animals using somatic cell nuclear transfer. Tabasco was used to produce the CHORI 242 BAC library which was used to generate the full pig genome sequence [226]. The clones were born by vaginal delivery and allowed to suckle. They were weaned at 4 weeks of age and continuously housed together. They were not vaccinated or ever in contact with other pigs after weaning. Pigs were fed once daily in the morning and had free access to water. Fecal samples were collected on day 14 (the last day of

that feed rotation) of each diet for a total of 4 samples for each animal. Samples were collected from the rectum into a sterile tube and frozen at -80 °C until time of analysis. Bovine rumen samples were collected as previously reported in ref.[218]. Chicken caeca were collected as previously reported in Qu *et al.* [162].

6.5.2 Sequencing

Swine and cattle samples were sequenced using PCR product from PCR specific primers flanking the V1-V3 region of bacterial 16S rDNA [227]. The forward fusion primers for pyrosequencing included 454 Life Science's A adapter, and barcode A fused to the 5' end of the V1 primer 27F. In chicken the V3 primer 341F was used. In all samples, the reverse fusion primer included 454 Life Science's B adapter (lowercase) fused to 5' end of V3 primer 534R. The fragments in the amplicon libraries were subjected to a single pyrosequence run from the V3 primer end using a 454 Life Science Genome Sequencer GS FLX (Roy J. Carver Biotechnology Center, University of Illinois).

6.5.3 Rank-abundance, Species-abundance, Preston Plots and Taxa Distributions

The reads from cattle and swine microbiomes were cleaned up using the method recommended in ref. [107]. For the chicken caecum microbiome we removed all sequences shorter than 100 bp. The ends of all reads were trimmed so that the sequences start and end in the same place in the 16S rRNA consensus structure. All remaining sequences were then aligned using the method described in ref. [216]. The OTUs were clustered using complete linkage [139] 3% sequence identity with the denominator 4 from [145] (counting indels as differences). The OTU abundance data for rank-abundance was then binned into a histogram using the method in Adami and Chu [228]. Species-abundance and Preston plots were generated following ref. [229]. Neutral model curves were generated using the algorithm for the sampling organisms from a neutral metacommunity [178]. Hubbell's θ parameter was fixed to match the exponentially decaying tail of the rank abundance. Offset was chosen by a least-squares method. Taxonomy assignments and comparison of libraries was made with the Library Compare tool [230] at RDP [16].

6.5.4 PCA Ordination

In Fig. 6.10 we show the results of Principal Component Analysis on our OTU data. In performing this calculation, each OTU was associated with a vector of real numbers of dimension $4L$ where L is the length of

the multiple alignment. The elements of the vectors were calculated in the following way. Each nucleotide within the multiple alignment is represented by a sub-vector of 4 numbers, A is $(1, 0, 0, 0)$, C is $(0, 1, 0, 0)$, G is $(0, 0, 1, 0)$, T is $(0, 0, 0, 1)$, whereas the gap is represented as $(0, 0, 0, 0)$. The vector associated with the OTU is then the arithmetic average of the vectors associated with each sequence within the OTU. We then perform the weighted PCA procedure [225] where we weigh each OTU by its abundance.

6.5.5 Closest-distance metric

We used the percent sequence distance metric in Fig. 6.14. The randomized dataset (red line) was generated in the following way. Each OTU (with its associated abundance) was replaced by a representative randomized sequence. This sequence was generated by selecting each nucleotide from a distribution of probabilities generated from the sequence reads. In this way, the base pair distribution for each position in the multiple alignment of the model dataset is the same as that of the experimental dataset. Furthermore, since the abundances of OTUs are kept, the rank abundance of the model dataset is exactly the same as that of the experimental dataset.

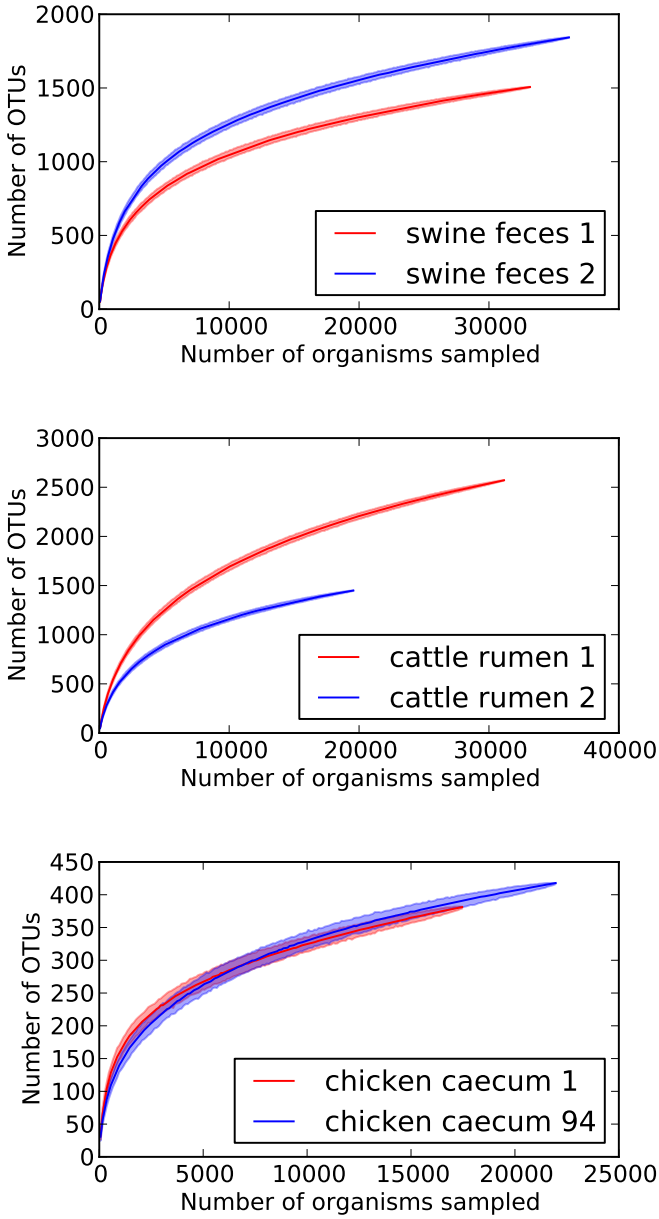


FIGURE 6.5: Rarefaction curves for the 6 vertebrate GI microbiomes.Solid line represents the median number of OTUs (100 resamplings) whereas the shaded area represents the 95% confidence interval.

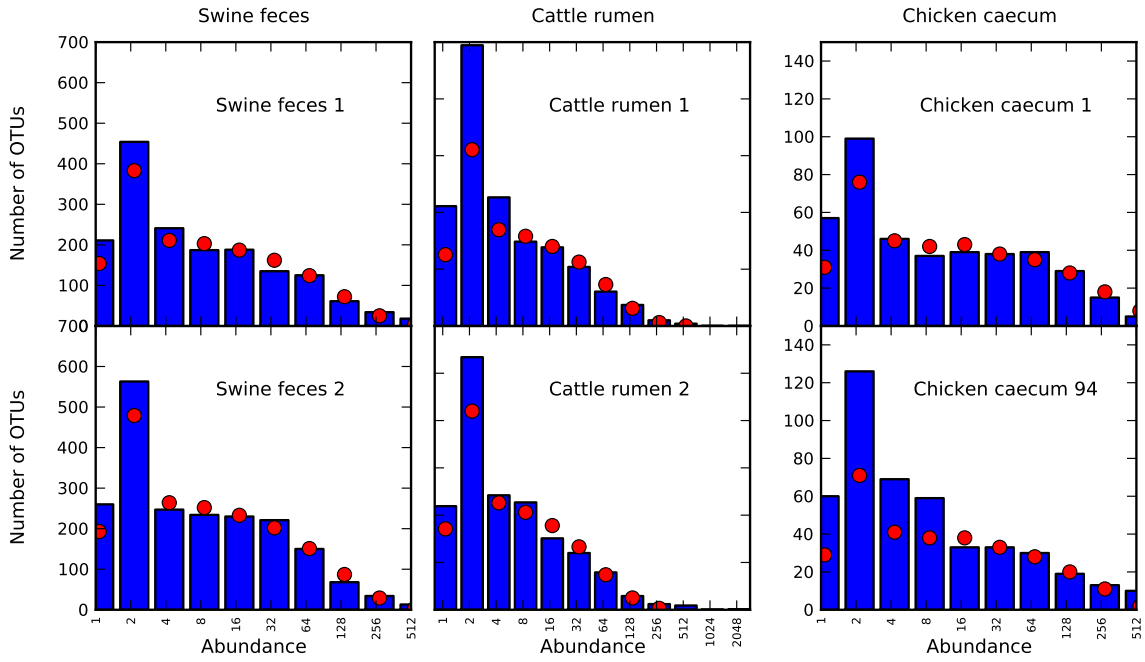


FIGURE 6.6: Preston plot for swine feces, cattle rumen and chicken caeca samples. In a Preston plot, the height of the bar indicates the number of species observed with abundance 1, 1-2, 2-4, 4-8, etc. Note that in all 6 datasets most OTUs are singletons. In this plot, 1-2 bars are highest because of an artifact. Traditionally, in a Preston plot, the OTUs with borderline abundances split evenly between two neighboring bins.

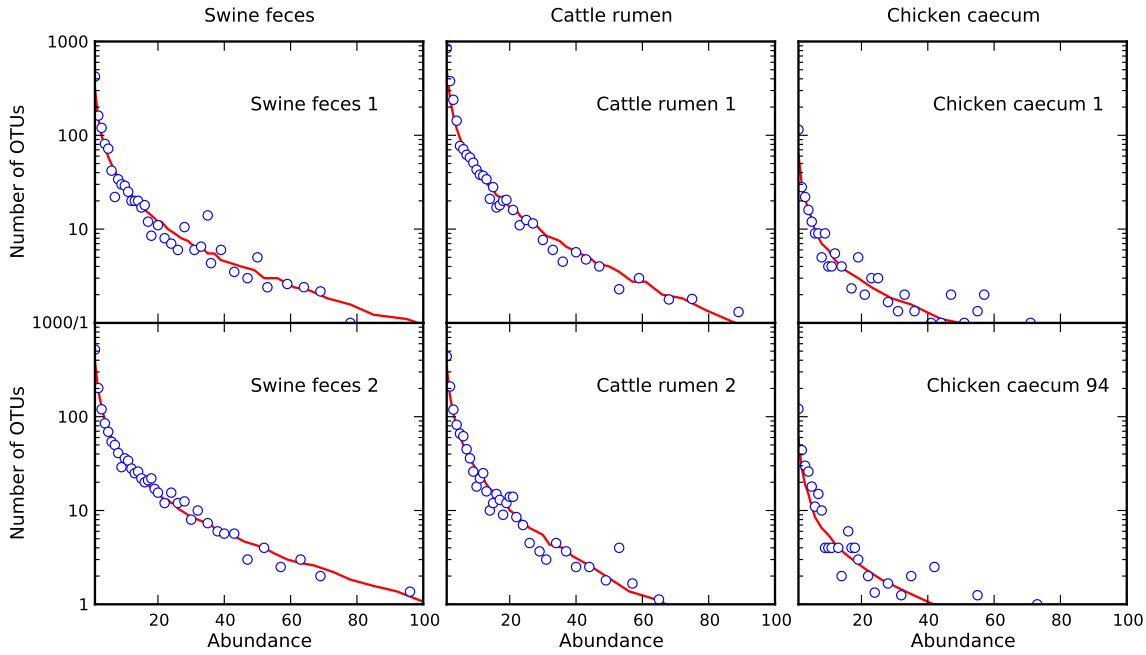


FIGURE 6.7: Species abundance distribution for swine feces, cattle rumen and chicken caeca. The species abundance distribution indicates the number of OTUs collected for each abundance.

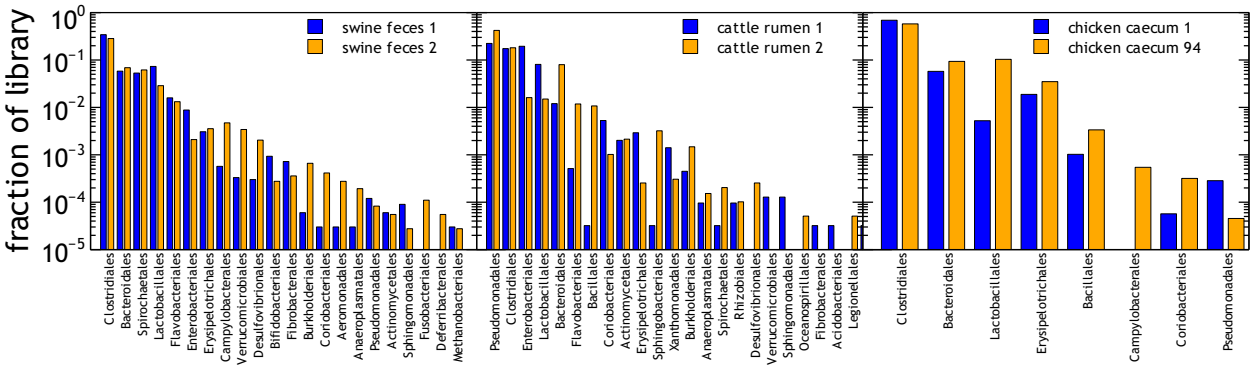


FIGURE 6.8: Taxa Comparisons. Taxonomic assignments at order level for all libraries, at 80% confidence threshold, sorted by combined abundance. Though there appear to be no differences in the form of the rank-abundance curves, we see differences in the taxonomic distributions here as the result of changes in diet or challenges to the microbial ecosystem.

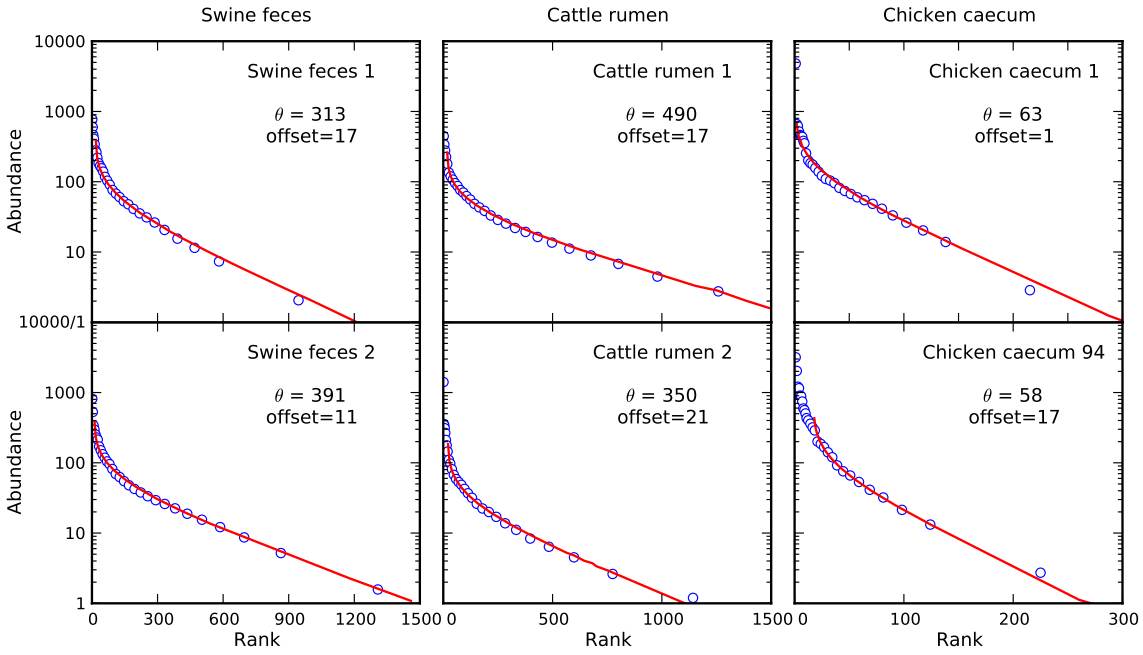


FIGURE 6.9: Comparison of rank abundance curves and neutral model fits for the six animal GI microbiomes. Lines indicate fits to the Hubbell's neutral metacommunity model. Parameter θ of the model is fit to correspond to the exponential tail in rank abundance. Offset represents the number of high-abundance OTUs that do not fit the neutral model.

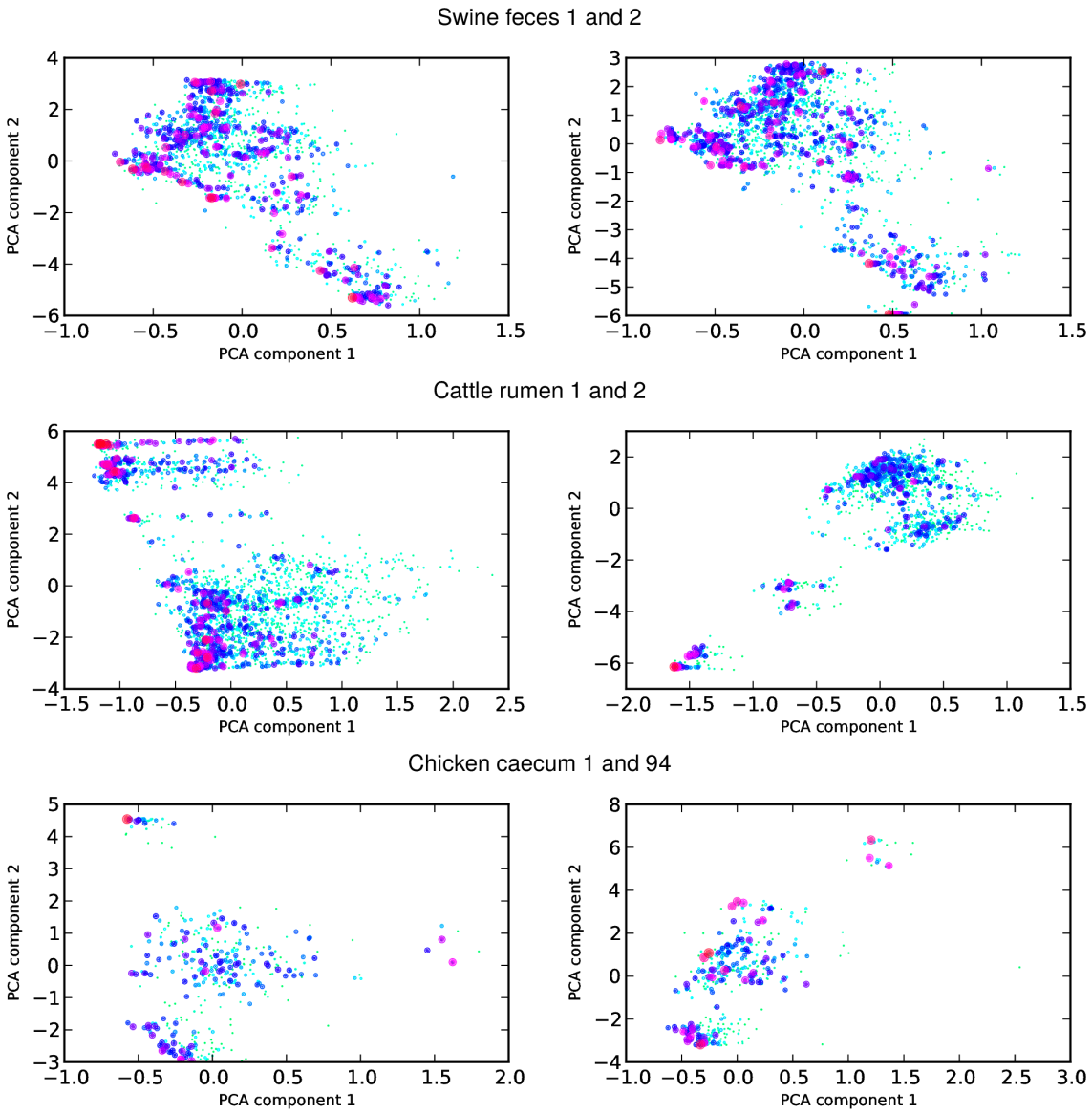


FIGURE 6.10: Weighted PCA ordination applied to the 6 experimental datasets. See the text for details on how weighted PCA was performed. Each circle in this Figure represents an OTU and its size and color indicates the logarithm of OTU abundance.

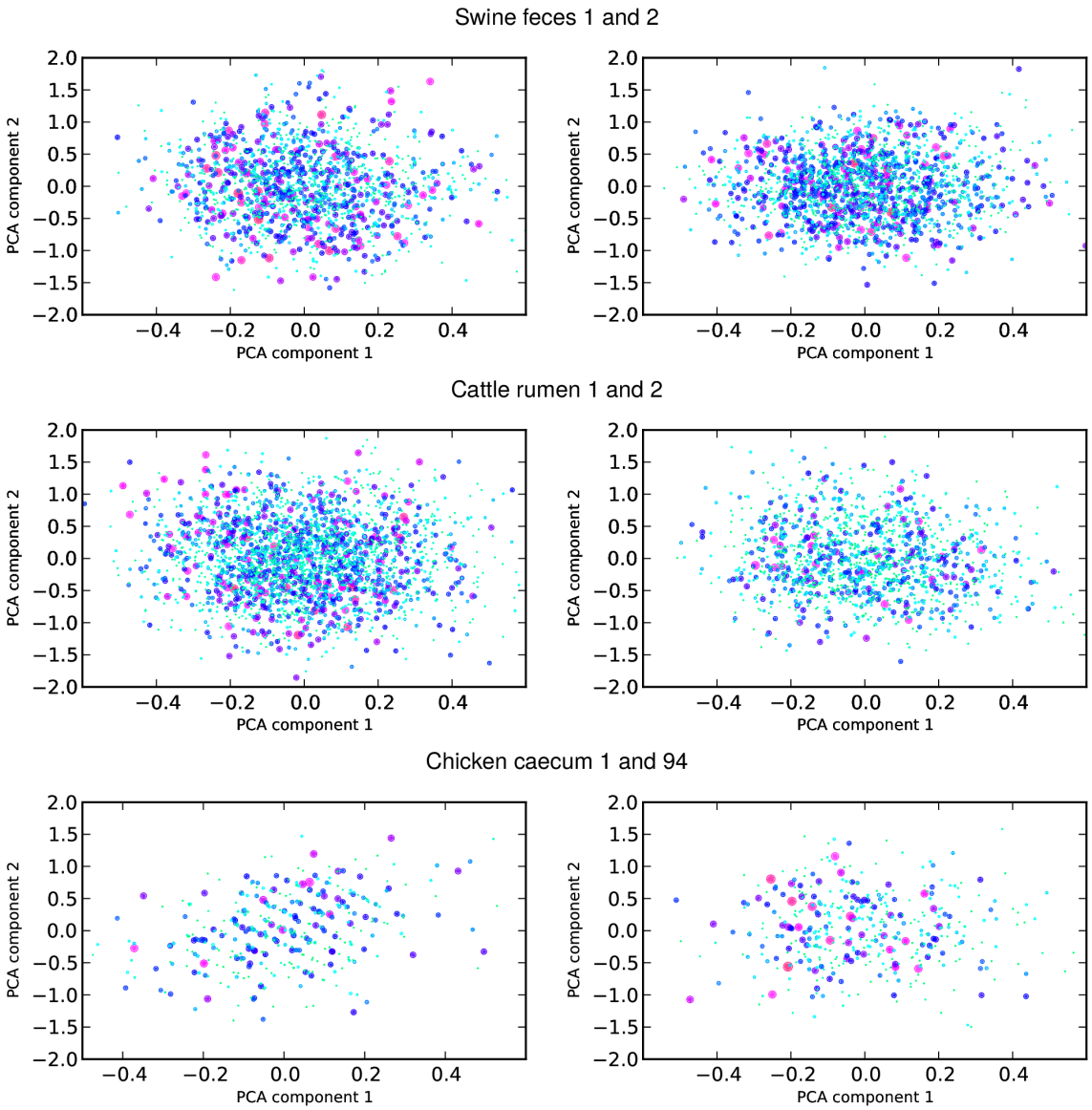


FIGURE 6.11: Weighted PCA ordination applied to the randomized datasets. Compare with Fig. 6.10. See the main text for details on how the randomized datasets were generated, and how weighted PCA was performed. Each circle in this Figure represents an OTU and its size and color indicates the logarithm of OTU abundance.

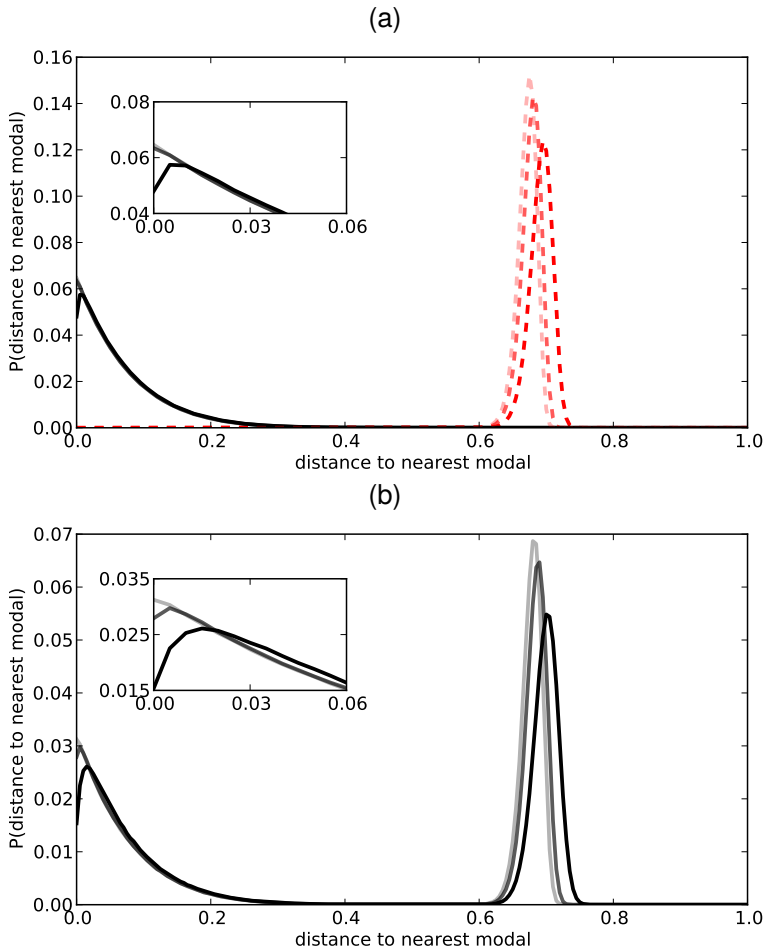


FIGURE 6.12: Measuring the effect of the choice of k on our metric. Darkest lines indicate $k = 2\%$, medium lines indicate $k = 6\%$ and lightest lines indicate $k = 10\%$. (a) $\alpha = 0.0$ model (red dashed lines) and $\alpha = 1.0$ model (black solid lines). (b) $\alpha = 0.5$ model (black solid lines).

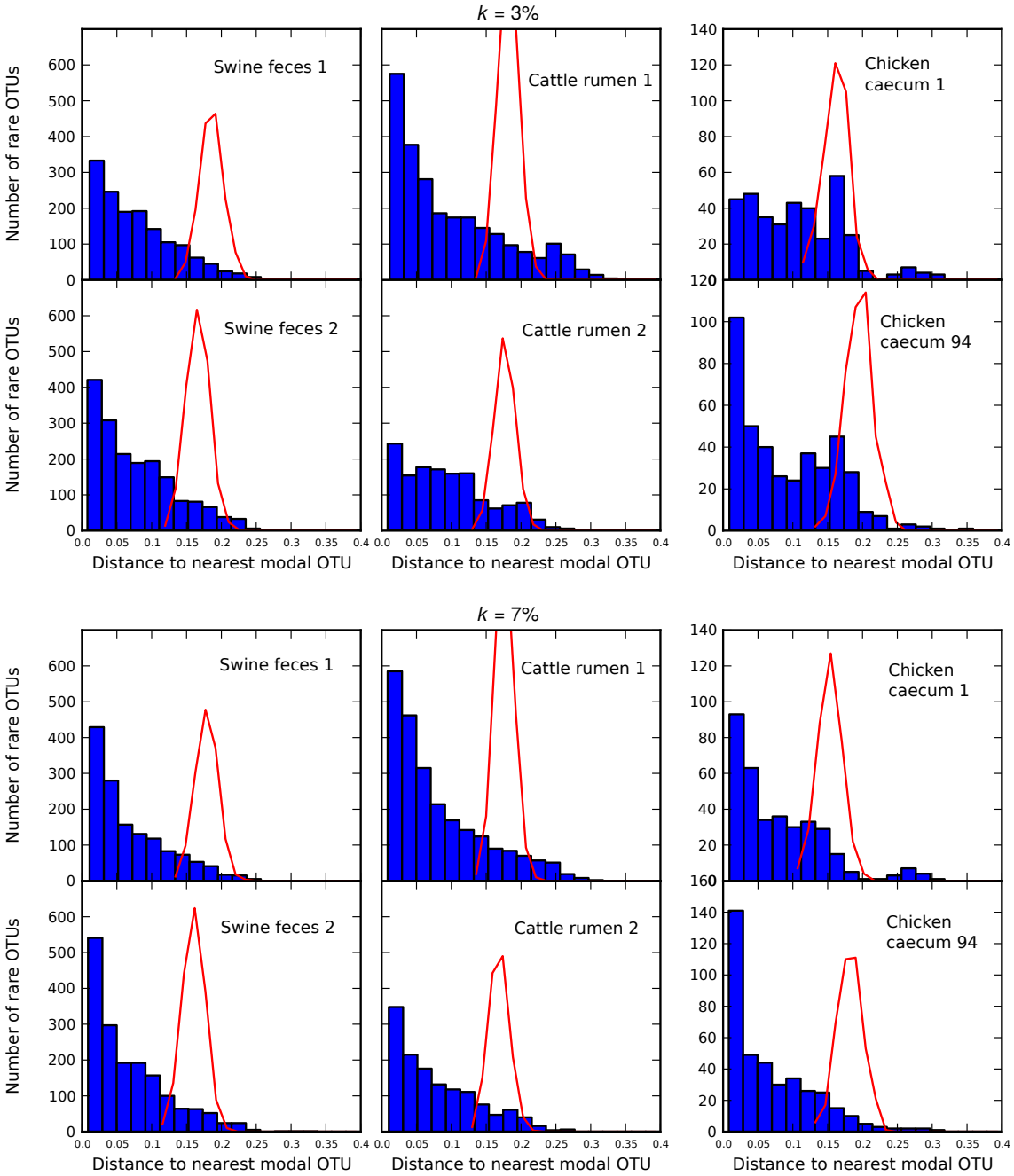


FIGURE 6.13: Histogram of distances of rare OTUs to the nearest modal OTU for each of the 6 GI microbiomes with cutoffs $k = 3\%$ and $k = 7\%$. Red dashed lines indicate the results of the metric applied to sequences that were randomized while preserving rank abundance and sequence statistics (see main text).

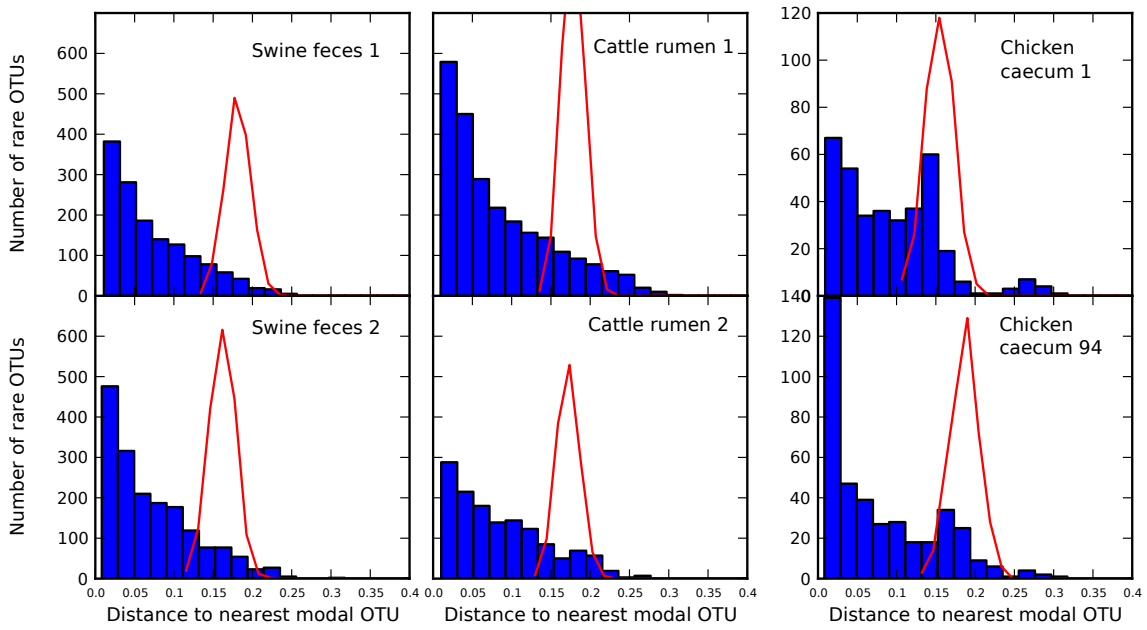


FIGURE 6.14: Histogram of distances of rare OTUs to the nearest modal OTU for each of the 6 gastrointestinal microbiomes with cutoff $k = 5\%$ (blue bars indicate experimental data). Red dashed lines indicate the results of the metric applied to sequences that were randomized while preserving rank abundance and sequence statistics (see text). Cattle and swine datasets share the same y-axis.

Chapter 7

Balance and structure in molecular phylogenies and taxonomies

The work in this Chapter came as a technical spin-off from the previous Chapter. For Chapter 6, a series of metrics were tried with the goal of being able to discriminate between various populations of interest found in microbial communities. This technique relies on balance measure of the subtree structure of the phylogenetic trees at all levels. Although unsuccessful in the original intent, I found that this technique could be used to ask more fundamental questions about the structure of phylogenies, and even compare them to other structures such as taxonomies. Although I will touch those questions briefly, this Chapter is otherwise strictly technical, and I will not try to answer those questions as this is work in progress. This Chapter is structured as follows. After introducing the technique, formulas and relevant terminology in Sec. 7.1, I explain the implementation and the data on which this technique is applied in Sec. 7.2, I show the resulting data in Sec. 7.4, and conclude with the future prospects in 7.4.

7.1 Introduction

As I have shown in the second part of this dissertation, there is a wealth of biology still waiting to be extracted from sequence and phylogenetic information that describe communities. The approach I will be using during this chapter started with the work of Herrada *et al.* [231]. They compared phylogenies of organisms in all domains of life, at different scales, by looking at the balance of these trees at many scales,

and after some data reduction they claimed to have found a universal, allometric scaling in a set of tree structure metrics. They also made similar claims for protein families [232]. Although the scaling claim has been strongly disputed [233], it opened the door to similar quantitative measures in trees.

The question of balance in life is not new in biology. Taxonomists and biologists, such as Ernst Mayr, claimed, to the point of dogmatism, that the structure of life must be balanced [234], and any systematization of biology must reflect this principle [235]. Since the dawn of the genomics era, numerous techniques have been created to infer the evolutionary relationships of organisms based on their genetic differences, and this "principle of balance" has been put into question [236]. The question has not been settled, and in the age of massive genetic data sets, there have been attempts to measure the true balance structure of evolutionary trees [237, 238].

The motivation of this work is to try to offer a way to compare the hand-based systematization present in taxonomies, versus the inferred results found in molecular phylogenies. This is work in progress, and what is presented here is most of the technical analysis necessary to make statements about the balance of these trees.

In what follows, I present the tree structure metrics used.

7.1.1 Tree structure metrics

We start by considering a tree structure. This tree contains a root node, leaf nodes and intermediate nodes. In an inferred phylogeny, the root is not necessarily well defined, but in a taxonomy a root always exist.

Now, let's consider a subtree S_i. This subtree is rooted at the i-th node, and includes itself and all its descendant nodes. Special cases are the whole tree (which is the largest possible subtree), and the leaf nodes, where each subtree contains only its root node.

For each subtree S_i, we can define its size A_i as the number of nodes in that subtree, including its root node.

A related quantity is the cumulative branch size C_i for the subtree S_i, which is the sum of branch sizes for all subtrees contained within subtree S_i, including itself, namely

$$C_i = \sum_{j \in S_i} A_j. \tag{7.1}$$

In principle, these quantities can be applied to any rooted tree structure. Of particular interest are binary

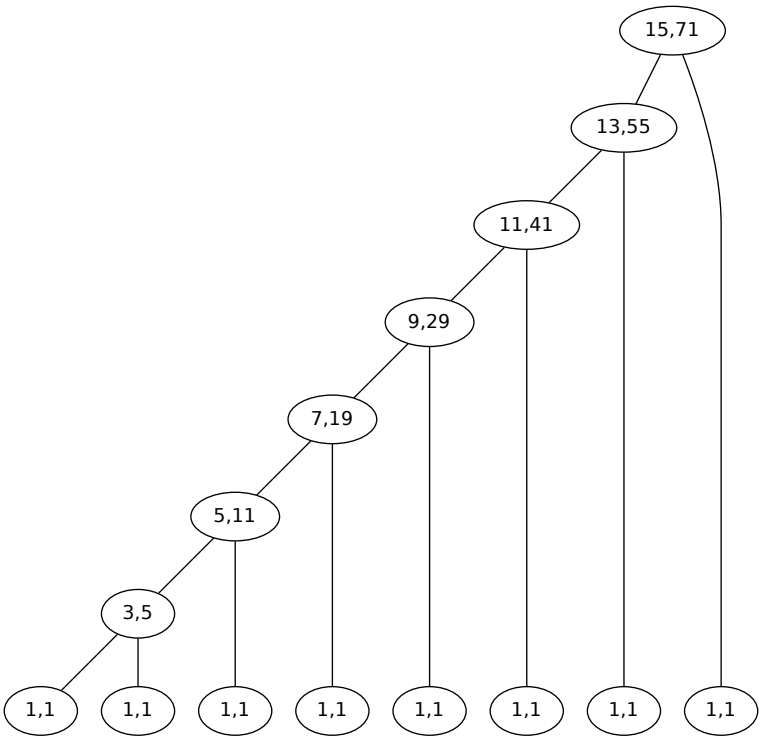


FIGURE 7.1: Example of A and C values calculated for a fully imbalanced binary tree with 8 leaf nodes.

trees, because of their usage in molecular phylogenies. There is a claim for a scaling relationship between C and A, $C \sim A^{\eta}$ [231].

Finally, a third related quantity can be defined using the previous two. The average node depth d_i of subtree S_i is the average distance of all leaf nodes to the root of the subtree, and it can be calculated from C_i and A_i as

$$d_i = \frac{C_i}{A_i} - 1 \tag{7.2}$$

For this last quantity, there are a series of limiting values of interest, which are defined for binary trees. First, for the case of a fully imbalanced binary tree (see Fig. 7.1 for an example), this quantity reaches its maximum, and as a function of A this can be written [232] as

$$d_{\max} = \frac{1}{4}\left(\frac{A^2 - 1}{A}\right) \tag{7.3}$$

For a fully balanced binary tree (see Fig. 7.2 for an example), the depth d reaches a minimum [232] for binary trees, scaling as

$$d_{\min} = \frac{1}{A}\left((A+1)\log_2(A+1) - 2A\right) \tag{7.4}$$

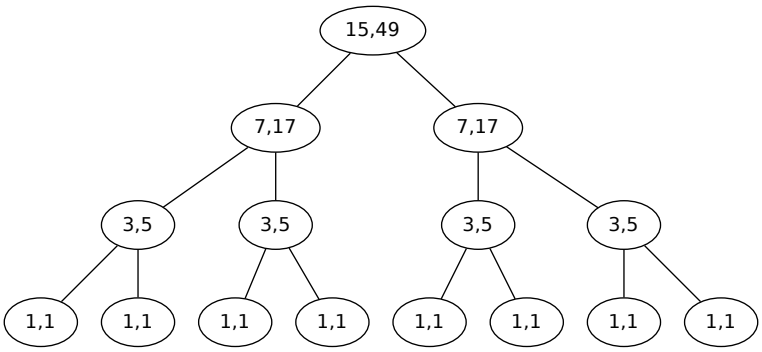


FIGURE 7.2: Example of A and C values calculated for a fully balanced binary tree with 8 leaf nodes.

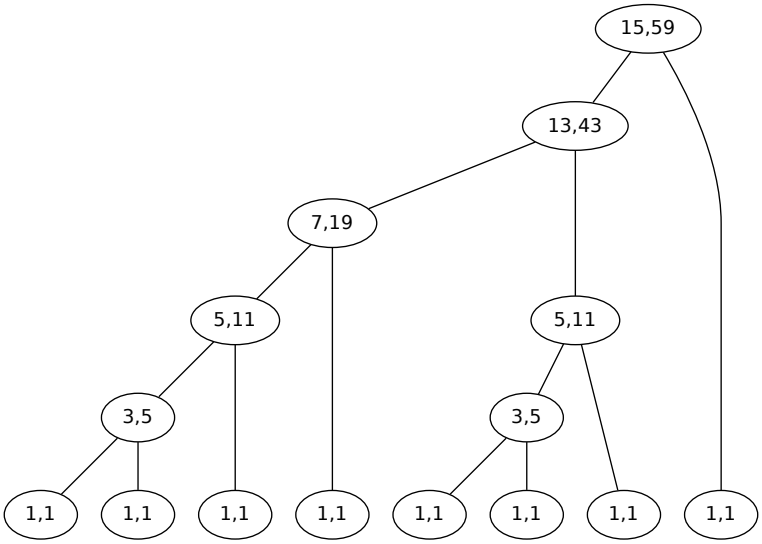


FIGURE 7.3: Example of A and C values calculated for a random binary tree structure with 8 leaf nodes.

Finally, if we move outside of the domain of binary trees, another interesting limit exists for a polytomic tree (see Fig. 7.4 for an example), where the scaling reaches an even lower value than eq. 7.4, and this scaling reads [231]

$$d_{\text{poly}} = 1 - \frac{1}{A} \tag{7.5}$$

All these quantities, A_i, C_i and d_i are calculated on a per-node basis, meaning that to be implemented numerically a program has to traverse the structure many times, although a dynamic programming approach would make this more efficient, especially for large enough trees.

In Figs. 7.1, 7.2, 7.3 and 7.4 I show examples of the different values of C_I and A_i for fully imbalanced, fully balanced, and random binary trees, as well as for a polytomic tree, respectively. The leaf nodes always have size 1, whereas the maximum values occur at the root. Overall they don't seem to be that different, but

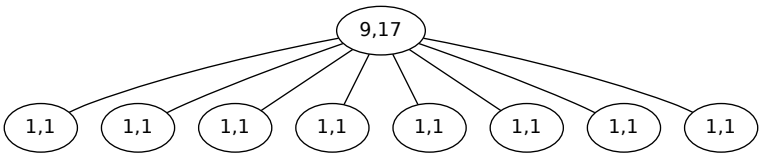


FIGURE 7.4: Example of A and C values calculated for a polytomic tree structure with 8 leaf nodes.

the values at the root of the trees say otherwise.

Now, for the purposes of this Chapter, I calculated these quantities for three large trees (hundreds of thousands of leaf nodes), and plotted the square root of the average node depth d_i as a function of subtree size A_i. The results are shown in Figs. 7.5, 7.6 and 7.5. The reason to plot it that way is because of the predicted behavior of d, scaling as $d \sim (\ln(A))^2$ for phylogenies [237, 239, 240]. Below, I describe the technical details of how these topological metrics of trees were calculated, and then present the results of the calculations.

7.2 Data sources and methods

7.2.1 16S rRNA phylogeny

The 16S rRNA phylogeny was created from a library of almost-full-length reads from the Greengenes database [15] (retrieved on November 16, 2011, containing 406,997 bacterial and archaeal sequences). The library was then aligned using the NAST algorithm [18], as implemented in the Mothur [139] software package, version 1.22, with the Greengenes template. After alignment, the Lane mask [113] was used to remove hard-to-align hypervariable regions of the 16S gene. This procedure does not seem to alter the large scale structure of a phylogeny created with this database, and it was designed to better resolve deep branching in 16S phylogenies. Finally, the alignment was used to construct an approximately Maximum Likelihood phylogeny using Fastree [143, 241] version 2.1.4, using parameters `-spr 4 -gamma -fastest -no2nd`, and no constraint trees.

7.2.2 Taxonomies

Two taxonomies were used in this chapter, the NCBI taxonomy [242, 243] (retrieved on November 16, 2011 and containing 744,131 leaf species) and the ENA-EMBL taxonomy [244] (retrieved on November 16, 2011 and containing 730,415 leaf species). Then I used the ETE [245] version 2.0 Python module to create a tree

version of the taxonomies that can be used to compare their structures to the 16S phylogeny.

7.2.3 Calculation of the structure quantities

I wrote a Python script that directly calculates from the tree structure of the 16S phylogenetic tree and the tree representation of the taxonomies the subtree size A_i, the cumulative branch size C_i and the average depth of nodes in the subtree of node i, d_i (as defined in eqs. 7.1 and 7.2). Since many nodes can generate the same values for A_i, C_i and d_i, redundant data was removed prior to plotting and further analysis, although in future analyses this data might be retained. No averaging or binning was performed on the resulting data.

7.3 Results

After generating the tree structures for the 16S library and the taxonomies, and then calculating the corresponding values for C_i, A_i and d_i, I plotted the data in various forms to understand their structure and make the comparisons between them. I did not perform any average or binning at this point because I wanted to see if there was any structure in the point clouds. I did remove redundant, non unique points generated by the many subtrees that share the same structure. At the very least, all leaf nodes will have the same values for A and C, as Figs. 7.2, 7.1, 7.3 and 7.4 show. This information might be relevant in future work done with this data.

Figure 7.5 shows the square root of average subtree depth d_i plotted as a function of the subtree size A_i for the Greengenes 16S phylogeny. The dashed, continuous and dotted curves show the limit cases of $d(A)$ for a fully imbalanced binary tree, fully balanced binary tree and polytomic tree. The plot show a great deal of structure, most of it falling within the limits for binary trees, leading to the interpretation that most of the subtrees lie somewhere in between being balanced and imbalanced. There are points falling outside of this binary region, which is surprising because reconstructed phylogenies are usually structured as binary trees, with branch lengths to indicate extra information about the nodes. Looking at the structure of these points, especially the ones that fall closer to the bottom, we see that the overall structure follows certain curves similar to the bottom polytomic limit, suggesting the presence of non-binary subtrees with varying degrees of balance. Upon closer inspection, there are even points falling on the polytomic curve, which is a rather surprising result considering that these measures do not account for branch lengths.

Figure 7.6 shows $d_i^{1/2}$ as a function of A_i for the two taxonomies considered, NCBI (red) and ENA-

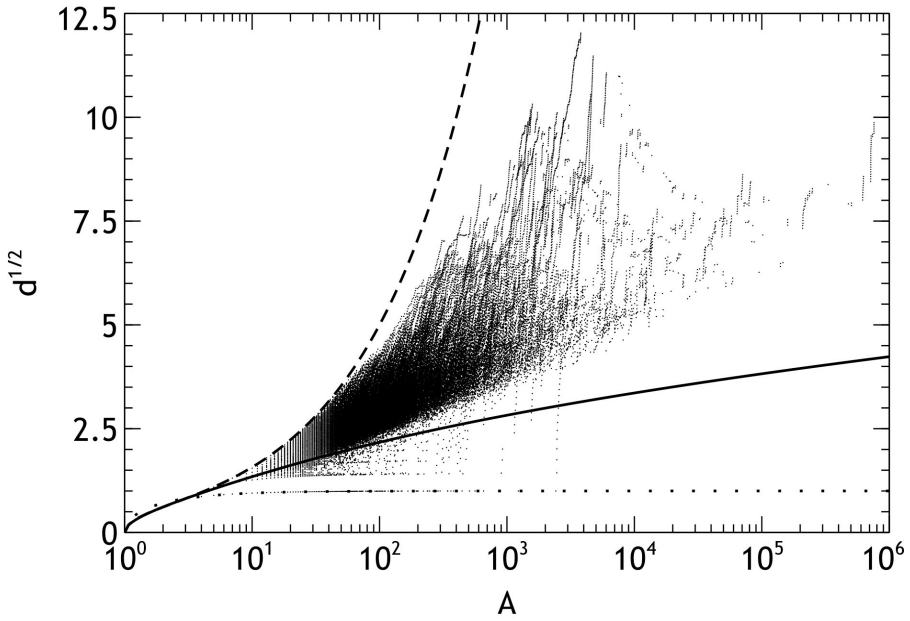


FIGURE 7.5: Square root of the average depth of nodes d versus subtree size A for the Greengenes phylogeny. The dashed, continuous and dotted curves represent the limit cases of $d(A)$ for fully imbalanced binary tree, fully balanced binary tree and polytomic tree, respectively. The points falling outside the limits for binary trees signal a deviation from the usually strict binary structure of software-reconstructed phylogenies, indicating either inference of non-binary clades or software artifacts.

EMBL (blue). The dashed, continuous and dotted curves have the same meaning described above. The point clouds are highly structured, and fall mostly outside the binary tree limits. This is not surprising, as relationships between taxa in these curated databases are not strictly binary. I can conjecture that the curves that these points seem to follow are the equivalent fully balanced limit curves for non binary trees, up to fully polytomic subtrees, thus making difficult to make an informed statement about the overall balance of these taxonomies. Also, both taxonomies show considerable overlap in their structures, as shown by the purple hues in the plot. This is expected as both databases are highly related and regularly exchange information.

Figure 7.7 shows a comparison between the structures of the 16S phylogeny (green) and the NCBI taxonomy (red), as seen from the values of d_i calculated for both. Clearly. both structures fall in different regions, with the 16S phylogeny being firmly in the binary tree region, and the NCBI away from the binary tree region. It is tempting to make a statement about balance of both structures, but we can only do so with the phylogeny, since most of it is represented as a binary structure. However the more complex structure of the taxonomy, with multiple kinds of subtree structures involved, makes this statement not possible at this moment. Finally, we can note a degree of overlap between some of the non-binary points of the phylogeny

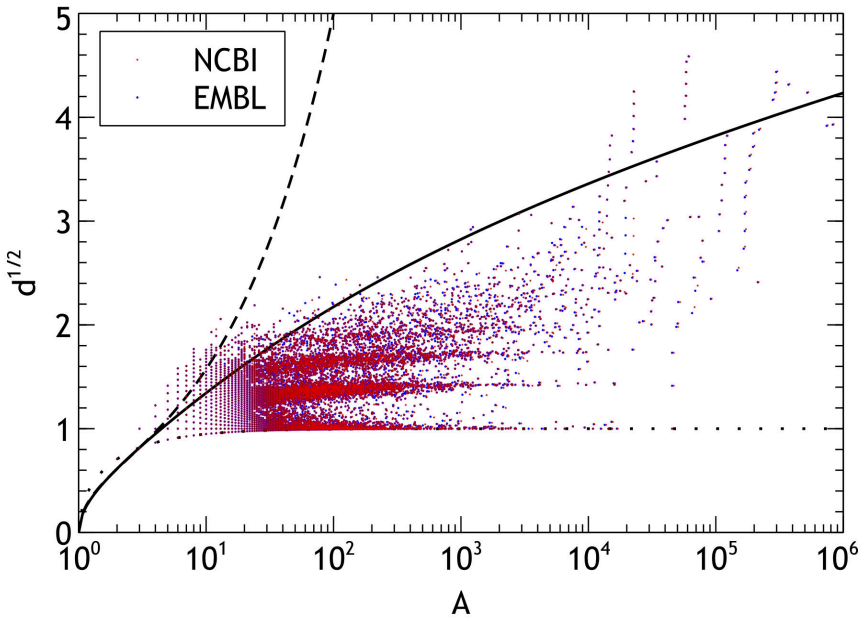


FIGURE 7.6: Square root of the average depth of nodes d versus subtree size A for the NCBI (red) and ENA-EMBL (blue) taxonomies. Purple coloring signals overlap between both taxonomies. The dashed, continuous and dotted curves represent the limit cases of $d(A)$ for fully imbalanced binary tree, fully balanced binary tree and polytomic tree, respectively. The plot shows that, unsurprisingly, the taxonomies mostly have a non-binary structure as the point cloud lies mostly outside the limits of the binary region. It also shows that both taxonomies mostly follow the same structure, with only a few differences.

in the lower part of the plot, and the points of the taxonomy. The structures roughly follow the same pattern, indicating the presence of similar subtree structures.

7.4 Discussion and conclusion

In this Chapter I calculated and explored metrics to probe the structure and balance of phylogenies and taxonomies. The metrics are based on the structures of all subtrees, and in particular we focused on the subtree size A_i, the cumulative branch size C_i and the average node depth d_i. We calculated these quantities for a phylogeny of 16S rRNA genes from the Greengenes database and for taxonomies obtained from NCBI and ENA-EMBL. We compared the resulting plots between them to highlight structural differences. I found that a reconstructed phylogeny like 16S and a hand-curated structure like the taxonomies studied have significantly different structure, with the phylogeny being primarily a binary tree (most likely by design), and the taxonomies having predominantly non-binary subtrees. If I can make an statement about the balance of the phylogeny, it is to note that it mostly does not lie close to the fully balanced binary case. This is

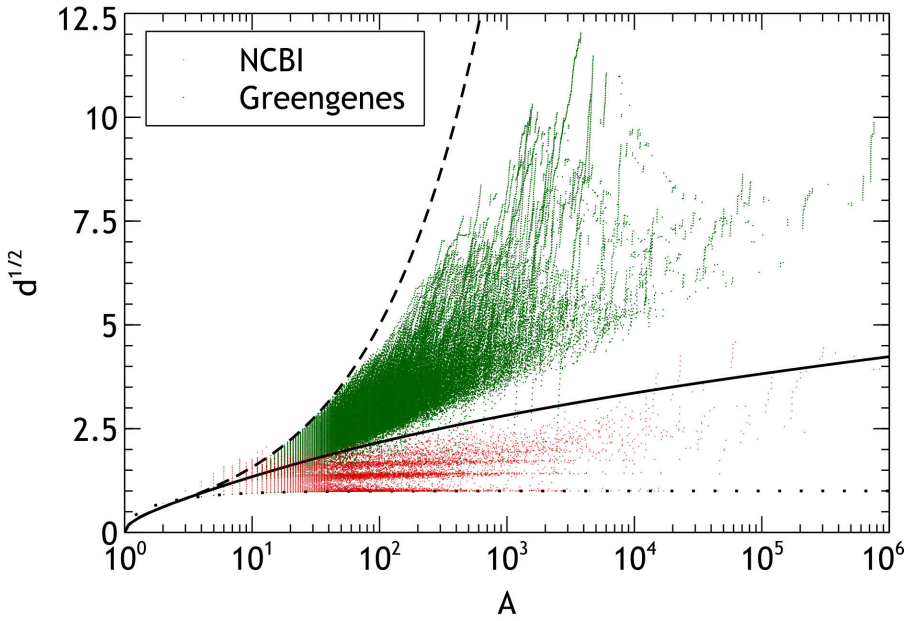


FIGURE 7.7: Square root of the average depth of nodes d versus subtree size A for the Greengenes phylogeny (green) and NCBI taxonomy (red). The dashed, continuous and dotted curves represent the limit cases of $d(A)$ for fully imbalanced binary tree, fully balanced binary tree and polytomic tree, respectively. This comparison highlights the major structural differences between the 16S phylogeny and the NCBI taxonomy.

relevant, as that limit also applies to a set of random binary trees used as test cases in phylogeny [231], which makes them less realistic when compared to inferred phylogenies.

While one of the motivations to start these calculations in the first place was to make a balance comparison between the phylogeny and the taxonomies, the current data makes this comparison impractical at the moment, as both structures are very different. This does not rule out that further work on this matter will eventually make the comparison possible. For instance, the phylogeny reconstructed here has more internal nodes and also contains branch length information, in lieu of the multi-branching subtrees present in taxonomy. This branch length information can be used, in principle, to infer the existence of these multi-branching structures in the phylogeny, especially in the case where the branch length is extremely small. Also, since I expect the existence of limit curves for the cases of subtrees that branch in modes other than binary, a way to map structures to a single master curve could be possible, a data collapse of sorts. More likely, a completely different way of assessing tree balance [246], or a combination of them, will allow fair comparisons between a molecular phylogeny and published taxonomies.

In light of this, I have not enough information at the moment to answer deeper questions related to

phylogenies and taxonomies, such as how well taxonomic structures and assumptions reflect the evolutionary history of organisms, the usefulness of taxonomies in light of very detailed, if seemingly unstructured, molecular phylogenies.

If I were to make a prediction about the answers, I would like to make speculation about what a more advanced analysis might reveal, borrowing an example from computer science. There is a now ubiquitous algorithm for information classification and retrieval, the PageRank algorithm [247], which yields rich yet seemingly unstructured results (much like a phylogeny), and was at the end vastly more powerful and successful than hand-curated indices of data (much like taxonomies) that were the basis of now forgotten web search engines. Considering that PageRank has been used beyond its original intended use [248], it is not far fetched to think that a similar situation might unfold in the realm of biology.

Chapter 8

Conclusion

In this dissertation, I devised computational approaches to model, analyze and understand two different stochastic systems in physics and biology.

In the first part, I constructed a Cell Dynamical Systems model of a quantum fluid. The model contains both a conservative and a dissipative part, and I incorporated a forcing term, allowing the driving of the system due to an external normal flow. The results show that it captures the essential physics of quantum fluids, exhibiting quantized vortices, and universal scaling laws in the distribution of defect velocities. This cellular, minimal approach made calculations more efficient allowing for detailed statistical analysis.

In the second part, I studied evolutionary dynamics of communities of microbes living in the gastrointestinal tract of vertebrates, from the pre-processing of the raw sequence data to the modeling of niche and neutral based evolution in sequence space.

I started by first quantifying how much phylogenetic information is lost when using partial instead of full sequence reads. Not surprisingly, the more initial data the better the recovery of the original phylogenetic structure. More surprising is the fact that the amount of information recovered is dependent of the region of the 16S gene used for these short reads. It's hard not to emphasize that care must be take when interpreting results from this type of data.

Then I proposed a pipeline that takes raw 16S short reads, removes known errors, aligns them and allows for curation. We show that a mixture of automatic, heuristic and manual approaches results in better alignments, diversity and phylogeny data than automatic processing alone.

Finally, using the tools above, I study the evolutionary dynamics of communities of bacteria living in

the gastrointestinal tracts of vertebrates. Starting from species abundance data, we show that communities seeming to evolve using neutral dynamics are shown to mostly obey niche-based dynamics when considering sequence data. Although the debate of niche vs. neutral dynamics is far from over, we show that modeling in sequence space can be a useful tool to probe evolutionary behavior using only sequence data.

8.1 Thoughts on interdisciplinary science

Since my late undergraduate years, I have been collaborating with biologists. First, with animal scientists and biochemists gaining insight on a seemingly simple process. Then, during my Doctorate, collaborating with microbiologists and bioinformaticians making sense of the processes involved in various microbial communities. And in the immediate future, I will be working with medical researchers, clinicians and system biologists. From a distance, it might look as if this was my intention all along, but uncertainties have plagued all the way.

I can only paraphrase this, but interdisciplinary science has been put together with sausage making and legislation in the sense that is better just to enjoy the end result and completely ignore the reality of the creation process, because it might disgust us [249]. Although an exaggeration, there is quite a bit of truth in that statement. During my years in the fuzzy border between fields, I have noticed very overt indifference, mistrust or even disdain between potential collaborators coming from different backgrounds. I guess this can be explained due to the natural fear of the unknown, but mostly I think the training, the upbringing of scientists has part to blame. Let me explain with my own training as an example.

I am trained as a theoretical physicist. We use advanced mathematics that are beyond the immediate understanding of most biologists without a few years of training. These techniques are mostly firmly rooted in physics, because they were devised as solution to problems in physics, and in my opinion learning them without understanding of their physical basis would diminish their power. Likewise, a basic knowledge of programming and computer science is very useful. I'm not asking for a full computer science degree and an intimate knowledge of how computers work, but an understanding on how to translate a potential solution or a method into a program that can implement those solutions. At the very least, this requires being comfortable with computers. This is not such a common trait, not even between physicists.

I cannot really speak for the pure biological training, but I have noticed a frequent aversion to mathematics and computer, much like physicists' stereotyped aversion towards chemistry. At most, math is the

last step in an experiment's protocol.

So, how do we bridge this gap? How is it possible to make science happen when mixing scientists with very different degrees of training? My take on the problem is the following. Besides a great deal of patience (and some diplomatic skills), I think the very first step is genuine interest in the problem at hand in the first place. Also, a willingness to learn the basics from the other field, even if uncomfortable at times, just like any learning process. But I also think that each party requires mastery of their own fields. Without it, it would be impossible to abstract the problem into terms and concepts we truly understand, and also because it makes the sharing of information possible. With this mastery, we can teach the very essentials to collaborators, for example from a simple log-log plot to the intricate heuristics involved in some bioinformatic software, without drowning them in ultimately unnecessary complexity.

A related question is, is it possible at all to train this hybrid scientist, without going through two Ph.D. degrees? I am not advocating for the training of a "Renaissance man" of sorts. I am also aware that in training this scientist we should be cautious about training someone with very broad, but shallow knowledge, a "jack of all trades, master of none." I am not sure about the answer. Today, traditional university departments are still very conservative about their curricula, and students must go out of their way to find "melting pots" with researchers willing to mentor an "outsider." Probably things will get clearer in the future, as fields grow out of their niches and mature. As for myself, I would not be who I am without the time spent working in superfluid simulations or year-long peer reviews in an unfamiliar setting. I am confident that the acquired skills will make my time easier when meeting and collaborating with yet another class of researchers. Exciting times lie ahead.

Appendix

Calculation of derivatives in cell dynamical systems models

We solve the time dependent real Ginzburg-Landau dynamics from Eq. 3.8 using the cell dynamical systems approach (CDS) [5, 6, 36]. Numerical efficiency becomes particularly important in 3D simulations and this method is tailored to that. The complex variable $\psi(\boldsymbol{r}, t)$ is replaced by a $\psi_{i,j}^{(n)}$ (or $\psi_{i,j,k}^{(n)}$ in 3D) defined on a square lattice of size $N \times N$ (a cube of size $N \times N \times N$ in 3D) at time n. The idea of the CDS method is to construct a discrete set of maps for each lattice cell such that the flow properties of the continuous dynamical system are preserved. A cell dynamics is defined by two steps: a local update

$$\tilde{\psi}^{(n+1)} = \frac{A\psi^{(n)}}{\sqrt{1 + \psi^{2(n)}(A^2 - 1)}}, \tag{A.1}$$

where $A > 1$ is a parameter that determines the global rate of convergence to the fixed points of the local double-well potential, and a global update taking into account the interactions between neighboring cells

$$\psi^{(n+1)} = \tilde{\psi}^{(n+1)} + C\nabla^2\tilde{\psi}^{(n+1)}, \tag{A.2}$$

where C is a constant proportional to the phenomenological diffusion constant. The isotropy of the order parameter being simulated naturally mandates the isotropy of the difference operators used to implement the coupled maps (see Tomita [38]). Oono and Puri [36] chose a 9 point stencil to implement a "Laplacian"

operator, which is highly isotropic for a 2D square lattice [39]. This stencil reads

$$\nabla^2\psi \equiv \frac{3}{dx^2}\left(\frac{1}{6}\sum_{\mathrm{NN}}\psi + \frac{1}{12}\sum_{\mathrm{NNN}}\psi - \psi\right), \tag{A.3}$$

where *NN* stands for the nearest neighbors in the discretized lattice and *NNN* are the next-to-nearest neighbors for each node in the lattice.

Considering the same isotropy requirements, the discretization of the 3D Laplace operator reads as [6]

$$\nabla^2 f \equiv \frac{3}{dx^2}\left(\frac{1}{9}\sum_{\mathrm{NN}}\psi + \frac{1}{36}\sum_{\mathrm{NNN}}\psi - \psi\right). \tag{A.4}$$

Since the calculation of the position and velocities of defects involves first order spatial derivatives of the ψ field, an isotropic discretization of the gradients is important in order to reduce the underlying lattice anisotropic effects. We use the isotropic version of the gradients both in 2D and 3D. Following the idea that these operators have to be accurately represented in Fourier space [39], we use the following stencil for 2D

$$\begin{aligned}\nabla_x\psi \;\equiv\; & \frac{1}{8dx}\Big(\psi_{i+1,j+1} + 2\psi_{i,j+1} - \psi_{i-1,j+1} + \\ & + \;\psi_{i+1,j-1} - 2\psi_{i,j-1} - \psi_{i-1,j-1}\Big),\end{aligned} \tag{A.5}$$

where i and j are the lattice indices for the x and y directions, respectively. Swapping indices we can obtain the corresponding expression for $\nabla_y\psi$. We note that this expression looks very similar to a 4 point first derivative in a 2D square lattice [250]. For the gradients in 3D, the stencil reads

$$\begin{aligned}\nabla_x\psi \;=\; & \frac{1}{8dx}\Big(\psi_{i+1,j+1,k} - \psi_{i-1,j+1,k} + \psi_{i+1,j-1,k} - \psi_{i-1,j-1,k} + \\ & + \;\psi_{i+1,j,k+1} - \psi_{i-1,j,k+1} + \psi_{i+1,j,k-1} - \psi_{i-1,j,k-1}\Big) + \\ & + \;\frac{1}{4dx}\Big(\psi_{i,j+1,k} - \psi_{i,j-1,k} + \psi_{i,j,k+1} - \psi_{i,j,k-1}\Big),\end{aligned} \tag{A.6}$$

with the corresponding index swap to obtain $\nabla_y\psi$ and $\nabla_z\psi$.

References

1. Pismen, L. *Vortices in nonlinear fields: From liquid crystals to superfluids, from non-equilibrium patterns to cosmic strings* (Clarendon Press, Oxford, 1999).

2. Pace, N. R., Sapp, J. & Goldenfeld, N. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proc. Natl. Acad. Sci. USA* **74,** 1–8 (2012).

3. Paoletti, M. & Lathrop, D. Quantum Turbulence. *Annu. Rev. Condens. Matter Phys.* **2,** 213–234 (2011).

4. Tsubota, M. & Kasamatsu, K. Quantized vortices and quantum turbulence. arXiv:1202.1863 (2012).

5. Mondello, M. & Goldenfeld, N. Scaling and vortex dynamics after the quench of a system with a continuous symmetry. *Phys. Rev. A* **42,** 5865–5872 (1990).

6. Mondello, M. & Goldenfeld, N. Scaling and vortex-string dynamics in a three-dimensional system with a continuous symmetry. *Phys. Rev. A* **45,** 657–664 (1992).

7. Nagaya, T., Orihara, H. & Ishibashi, Y. Coarsening Dynamics of +1 and -1 Disclinations in Two-Dimensionally Aligned Nematics –Spatial Distribution of Disclinations. *J. Phys. Soc. Jpn.* **64,** 78–85 (1995).

8. Goldenfeld, N., Chan, P. Y. & Veysey, J. Dynamics of precipitation pattern formation at geothermal hot springs. *Phys. Rev. Lett.* **96,** 254501 (2006).

9. Ginzburg, V. L. & Pitaevskii, L. On the theory of superfluidity. *Sov. Phys. JETP* **34,** 858–861 (1958).

10. Ginzburg, V. L. & Sobyanin, A. A. Superfluidity of helium II near the λ point. *Sov. Phys. Uspekhi* **31,** 289–299 (1988).

11. Halperin, B. in *Physics of Defects, Proceedings of the Les Houches Summer School, Session XXXV* (eds Balian, R., Kleman, M. & Poirier, J.-P.) (North-Holland, Amsterdam, 1981), 814–857.

12. Mazenko, G. Velocity distribution for strings in phase-ordering kinetics. *Phys. Rev. E* **59,** 1574–1584 (1999).

13. Succi, S. *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond* (Clarendon Press, Oxford, 2001).

14. Chen, S. & Doolen, G. D. Lattice Boltzmann Method For Fluid Flows. *Annu. Rev. Fluid Mech.* **30,** 329–364 (1998).

15. DeSantis, T. Z. *et al.* NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* **34,** W394–W399 (2006).

16. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37,** D141–D145 (2009).

17. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25,** 1335–1337 (2009).

18. DeSantis, T. Z. *et al.* Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).

19. Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Math. Biosci.* **53,** 131–147 (1981).

20. *Quantized Vortex Dynamics and Superfluid Turbulence* (eds Barenghi, C. F., Donnelly, R. J. & Vinen, W. F.) (Springer, Berlin, 2001).

21. Lifshitz, E. M. & Landau, L. D. *Fluid Mechanics, Second Edition: Volume 6 (Course of Theoretical Physics)* (Butterworth-Heinemann, Oxford, 1987).

22. Coste, C. Nonlinear Schrödinger equation and superfluid hydrodynamics. *Eur. Phys. J. B* **1,** 245–253 (1998).

23. Schwarz, K. W. Generation of Superfluid Turbulence Deduced from Simple Dynamical Rules. *Phys. Rev. Lett.* **49,** 283–285 (1982).

24. Vinen, W. An Introduction to Quantum Turbulence. *J. Low Temp. Phys.* **145,** 7–24 (2006).

25. Baggaley, A. W. The Sensitivity of the Vortex Filament Method to Different Reconnection Models. *J. Low Temp. Phys.* 1–13 (2012).

26. Henderson, K. L. & Barenghi, C. F. Transition from Ekman flow to Taylor vortex flow in superfluid helium. *J. Fluid Mech.* **508,** 319–331 (2004).

27. Khalatnikov, I. M. in *An introduction to the theory of superfluidity* (ed Pines, D.) 105–110 (W.A. Benjamin, Inc., New York, 1965).

28. Aranson, I. & Steinberg, V. Spin-up and nucleation of vortices in superfluid ^{4}He. *Phys. Rev. B* **54,** 13072–13082 (1996).

29. Aranson, I. S., Kopnin, N. B. & Vinokur, V. M. Nucleation of Vortices by Rapid Thermal Quench. *Phys. Rev. Lett.* **83,** 2600–2603 (1999).

30. Aranson, I. S., Kopnin, N. B. & Vinokur, V. M. Dynamics of vortex nucleation by rapid thermal quench. *Phys. Rev. B* **63,** 184501 (2001).

31. Geurst, J. A. General theory unifying and extending the Landau-Khalatnikov, Ginzburg-Pitaevskii, and Hills-Roberts theories of superfluid ^{4}He. *Phys. Rev. B* **22,** 3207–3220 (1980).

32. Carlson, N. N. A topological defect model of superfluid vortices. *Physica D* **98,** 183–200 (1996).

33. Oono, Y. & Puri, S. Computationally efficient modeling of ordering of quenched phases. *Phys. Rev. Lett.* **58,** 836–839 (1987).

34. Veysey, J. & Goldenfeld, N. Watching rocks grow. *Nat. Phys.* **4,** 310–313 (2008).

35. Zapotocky, M., Goldbart, P. M. & Goldenfeld, N. Kinetics of phase ordering in uniaxial and biaxial nematic films. *Phys. Rev. E* **51,** 1216–1235 (1995).

36. Oono, Y. & Puri, S. Study of phase-separation dynamics by use of cell dynamical systems. I. Modeling. *Phys. Rev. A* **38,** 434–453 (1988).

37. Bahiana, M. & Massunaga, M. S. O. Cell-dynamics modeling of oscillator systems. *Phys. Rev. E* **52,** 321–326 (1995).

38. Tomita, H. Preservation of isotropy at the mesoscopic stage of phase separation processes. *Prog. Theor. Phys.* **85,** 47–56 (1991).

39. Teixeira, P. I. C. & Mulder, B. M. Comment on "Study of phase-separation dynamics by use of cell dynamical systems. I. Modeling". *Phys. Rev. E* **55,** 3789–3791 (1997).

40. Puri, S. & Oono, Y. Study of phase-separation dynamics by use of cell dynamical systems. II. Two-dimensional demonstrations. *Phys. Rev. A* **38,** 1542–1565 (1988).

41. Denniston, C., Marenduzzo, D., Orlandini, E. & Yeomans, J. M. Lattice Boltzmann algorithm for three-dimensional liquid-crystal hydrodynamics. *Philos. T. R. Soc. A* **362,** 1745–1754 (2004).

42. Leadbeater, M., Winiecki, T., Samuels, D. C., Barenghi, C. F. & Adams, C. S. Sound Emission due to Superfluid Vortex Reconnections. *Phys. Rev. Lett.* **86,** 1410–1413 (2001).

43. Bewley, G. P., Paoletti, M. S., Sreenivasan, K. R. & Lathrop, D. P. Characterization of reconnecting vortices in superfluid helium. *Proc. Natl. Sci. Acad. USA* **105,** 13707–13710 (2008).

44. Rocha, J. V. Scaling Solution for Small Cosmic String Loops. *Phys. Rev. Lett.* **100,** 071601 (2008).

45. Guttenberg, N. & Goldenfeld, N. Ordering dynamics in type-II superconductors. *Phys. Rev. E* **74,** 066202 (2006).

46. Ostermeier, R. M. & Glaberson, W. I. Instability of vortex lines in the presence of axial normal fluid flow. *J. Low Temp. Phys.* **21,** 191–196 (1975).

47. Samuels, D. C. Response of superfluid vortex filaments to concentrated normal-fluid vorticity. *Phys. Rev. B* **47,** 1107–1110 (1993).

48. Dombre, T. *et al.* Chaotic streamlines in the ABC flows. *J. Fluid Mech.* **167,** 353–391 (1986).

49. Barenghi, C. F., Samuels, D. C., Bauer, G. H. & Donnelly, R. J. Superfluid vortex lines in a model of turbulent flow. *Phys. Fluids* **9,** 2631–2643 (1997).

50. Perez, S. E., Hibberd, K., Stone, M. & Visser, M. Wave equation for sound in fluids with vorticity. *Physica D* **191,** 121–136 (2004).

51. Smith, M. R., Donnelly, R. J., Goldenfeld, N. & Vinen, W. F. Decay of vorticity in homogeneous turbulence. *Phys. Rev. Lett.* **71,** 2583–2586 (1993).

52. Reisenegger, A. The spin-up problem in Helium II. *J. Low Temp. Phys.* **92,** 77–106 (1993).

53. Warszawski, L., Melatos, A. & Berloff, N. Unpinning triggers for superfluid vortex avalanches. *Phys. Rev. B* **85,** 104503 (2012).

54. Bray, A. & Humayun, K. Universal amplitudes of power-law tails in the asymptotic structure factor of systems with topological defects. *Phys. Rev. E* **47,** R9–R12 (1993).

55. Qian, H. & Mazenko, G. Vortex dynamics in a coarsening two-dimensional XY model. *Phys. Rev. E* **68,** 021109 (2003).

56. Huepe, C., Riecke, H., Daniels, K. & Bodenschatz, E. Statistics of defect trajectories in spatio-temporal chaos in inclined layer convection and the complex Ginzburg–Landau equation. *Chaos* **14,** 864–874 (2004).

57. Mazenko, G. Vortex velocities in the O (n) symmetric time-dependent Ginzburg-Landau model. *Phys. Rev. Lett.* **78,** 401–404 (1997).

58. Mazenko, G. Defect statistics in the two-dimensional complex Ginzburg-Landau model. *Phys. Rev. E* **64,** 016110 (2001).

59. Qian, H. & Mazenko, G. Growth of order in an anisotropic Swift-Hohenberg model. *Phys. Rev. E* **73,** 036117 (2006).

60. Angheluta, L., Jeraldo, P. & Goldenfeld, N. Anisotropic velocity statistics of topological defects under shear flow. *Phys. Rev. E* **85,** 011153 (2012).

61. Chaté, H. & Manneville, P. Phase diagram of the two-dimensional complex Ginzburg-Landau equation. *Physica A* **224,** 348–368 (1996).

62. Aranson, I. & Kramer, L. The world of the complex Ginzburg-Landau equation. *Rev. Mod. Phys.* **74,** 99–143 (2002).

63. Bray, A. Theory of phase-ordering kinetics. *Adv. Phys.* **51,** 481–587 (1994).

64. Bodenschatz, E., Pesch, W. & Ahlers, G. Recent developments in Rayleigh-Bénard convection. *Annu. Rev. Fluid Mech.* **32,** 709–778 (2000).

65. Groma, I. & Bakó, B. Probability distribution of internal stresses in parallel straight dislocation systems. *Phys. Rev. B* **58,** 2969–2974 (1998).

66. Miguel, M., Vespignani, A., Zapperi, S., Weiss, J. & Grasso, J. Intermittent dislocation flow in viscoplastic deformation. *Nature* **410,** 667–671 (2001).

67. Paoletti, M., Fisher, M., Sreenivasan, K. & Lathrop, D. Velocity statistics distinguish quantum turbulence from classical turbulence. *Phys. Rev. Lett.* **101,** 154501 (2008).

68. White, A., Barenghi, C., Proukakis, N., Youd, A. & Wacks, D. Nonclassical velocity statistics in a turbulent atomic Bose-Einstein condensate. *Phys. Rev. Lett.* **104,** 75301 (2010).

69. Adachi, H. & Tsubota, M. Numerical study of velocity statistics in steady counterflow quantum turbulence. *Phys. Rev. B* **83,** 132503 (2011).

70. Ispánovity, P., Groma, I., Gyorgyi, G., Csikor, F. & Weygand, D. Submicron plasticity: yield stress, dislocation avalanches, and velocity distribution. *Phys. Rev. Lett.* **105,** 85503 (2010).

71. Daniels, K. & Bodenschatz, E. Statistics of defect motion in spatiotemporal chaos in inclined layer convection. *Chaos* **13,** 55–63 (2003).

72. Chavanis, P. & Sire, C. Statistics of velocity fluctuations arising from a random distribution of point vortices: The speed of fluctuations and the diffusion coefficient. *Phys. Rev. E* **62,** 490–506 (2000).

73. Bray, A. Velocity distribution of topological defects in phase-ordering systems. *Phys. Rev. E* **55,** 5297–5301 (1997).

74. Ispánovity, P., Groma, I., Györgyi, G., Szabó, P. & Hoffelner, W. Criticality of Relaxation in Dislocation Systems. *Phys. Rev. Lett.* **107,** 85506 (2011).

75. Boyer, D. & Viñals, J. Grain boundary pinning and glassy dynamics in stripe phases. *Phys. Rev. E* **65,** 046119 (2002).

76. Boyer, D. Numerical study of domain coarsening in anisotropic stripe patterns. *Phys. Rev. E* **69,** 066111 (2004).

77. Pesch, W. & Behn, U. in *Evolution of Spontaneous Structures in Dissipative Continuous Systems* (eds Busse, F. & Müller, S.) 335–383 (Springer, Berlin, 1998).

78. Daniels, K. & Bodenschatz, E. Defect turbulence in inclined layer convection. *Phys. Rev. Lett.* **88,** 34501 (2002).

79. Greenside, H., Cross, M. & Coughran, J. W. Mean flows and the onset of chaos in large-cell convection. *Phys. Rev. Lett.* **60,** 2269–2272 (1988).

80. Rehberg, I., Rasenat, S. & Steinberg, V. Traveling waves and defect-initiated turbulence in electro-convecting nematics. *Phys. Rev. Lett.* **62,** 756–759 (1989).

81. Beta, C., Mikhailov, A., Rotermund, H. & Ertl, G. Defect-mediated turbulence in a catalytic surface reaction. *Europhys. Lett.* **75,** 868–874 (2006).

82. Gil, L., Lega, J., Meunier, J., *et al.* Statistical properties of defect-mediated turbulence. *Phys. Rev. A* **41,** 1138–1141 (1990).

83. Kaiser, M., Pesch, W. & Bodenschatz, E. Mean flow effects in the electro-hydrodynamic convection in nematic liquid crystals. *Physica D* **59,** 320–333 (1992).

84. Hildebrand, M., Bär, M. & Eiswirth, M. Statistics of topological defects and spatiotemporal chaos in a reaction-diffusion system. *Phys. Rev. Lett.* **75,** 1503–1506 (1995).

85. Song, K., Sun, Z. & An, L. Defect evolution and hydrodynamic effects in lamellar ordering process of two-dimensional quenched block copolymers. *J. Chem. Phys.* **130,** 124907 (2009).

86. Kumaran, V. & Raman, D. Shear alignment of a disordered lamellar mesophase. *Phys. Rev. E* **83,** 031501 (2011).

87. Zilman, A. & Granek, R. Undulation instability of lamellar phases under shear: A mechanism for onion formation? *Eur. Phys. J. B* **11,** 593–608 (1999).

88. Zapotocky, M., Goldbart, P. & Goldenfeld, N. Kinetics of phase ordering in uniaxial and biaxial nematic films. *Phys. Rev. E* **51,** 1216–1235 (1995).

89. Christensen, J. & Bray, A. Pattern dynamics of Rayleigh-Bénard convective rolls and weakly segregated diblock copolymers. *Phys. Rev. E* **58,** 5364–5370 (1998).

90. Min, I., Mezić, I. & Leonard, A. Levy stable distributions for velocity and velocity difference in systems of vortex elements. *Phys. Fluids* **8,** 1169–1180 (1996).

91. Shinozaki, A. & Oono, Y. Spinodal decomposition in 3-space. *Phys. Rev. E* **48,** 2622–2654 (4 1993).

92. Pismen, L. Mean flow effects in defects dynamics. *Physica D* **61,** 217–226 (1992).

93. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).

94. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. USA* **105,** 3805–3810 (2008).

95. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA* **103,** 12115–12120 (2006).

96. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).

97. McKenna, P. *et al.* The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog.* **4,** e20 (2008).

98. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).

99. Brazelton, W. J. *et al.* Archaea and bacteria with surprising microdiversity show shifts in dominance over 1,000-year time scales in hydrothermal chimneys. *Proc. Natl. Acad. Sci. USA* **107,** 1612–1617 (2010).

100. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311,** 496–503 (2006).

101. Elshahed, M. S. *et al.* Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl. Environ. Microbiol.* **74,** 5422–5428 (2008).

102. Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D. & Knight, R. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.* **35,** e120 (2007).

103. Liu, Z., DeSantis, T. Z., Andersen, G. L. & Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* **36,** e120 (2008).

104. Huse, S. M. *et al.* Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* **4,** e1000255 (2008).

105. Youssef, N. *et al.* Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl. Environ. Microbiol.* **75,** 5227–5236 (2009).

106. Quince, C. *et al.* Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* **6,** 639–641 (2009).

107. Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12,** 118–123 (2010).

108. Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* **3,** 1314–1317 (2009).

109. Farris, J. S. The Meaning of Relationship and Taxonomic Procedure. *Syst. Zool.* **16,** 44 –51 (1967).

110. Farahi, K., Pusch, G. D., Overbeek, R. & Whitman, W. B. Detection of lateral gene transfer events in the prokaryotic tRNA synthetases by the ratios of evolutionary distances method. *J. Mol. Evol.* **58,** 615–631 (2004).

111. Robinson, D. & Foulds, L. in *Combinatorial Mathematics IV* (eds Horadam, A. F. & Wallis, W. D.) **748** (Springer-Verlag, Berlin, 1979), 119–126.

112. Phipps, J. B. Dendrogram topology. *Syst. Zool.* **20,** 306–308 (1971).

113. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. USA* **82,** 6955–6959 (1985).

114. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).

115. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).

116. Ott, M., Zola, J., Stamatakis, A. & Aluru, S. in *Proceedings of the 2007 ACM/IEEE conference on Supercomputing - SC '07* (ACM Press, New York, 2007), 1–11.

117. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* **57,** 758–771 (2008).

118. Stamatakis, A. in *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium* (IEEE, 2006), 1–8.

119. Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E. & Stamatakis, A. How many bootstrap replicates are necessary? *J. Comput. Biol.* **17,** 337–354 (2010).

120. Bininda-Emonds, O. R., Brady, S. G., Kim, J. & Sanderson, M. J. in *Pacific Symposium on Biocomputing* (2001), 547–558.

121. Moret, B., Roshan, U. & Warnow, T. in *WABI 2002* (eds Guigó, R. & Gusfield, D.) **2452** (Springer-Verlag, Berlin, 2002), 343–356.

122. Shi, Y., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459,** 266–269 (2009).

123. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).

124. Turnbaugh, P. J. *et al.* Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl. Acad. Sci. USA* **107,** 7503–7508 (2010).

125. Jones, R. T., Knight, R. & Martin, A. P. Bacterial communities of disease vectors sampled across time, space, and species. *ISME J.* **4,** 223–231 (2010).

126. Koopman, M. M., Fuselier, D. M., Hird, S. & Carstens, B. C. The carnivorous pale pitcher plant harbors diverse, distinct, and time-dependent bacterial communities. *Appl. Environ. Microbiol.* **76,** 1851–1860 (2010).

127. Lozupone, C. A., Hamady, M. & Knight, R. UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7,** 371 (2006).

128. Schloss, P. D. The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. *PLoS Comp. Biol.* **6,** e1000844 (2010).

129. Frigaard, N.-U., Martinez, A., Mincer, T. J. & DeLong, E. F. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439,** 847–850 (2006).

130. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: A ribosomal RNA approach. *Annu. Rev. Microbiol.* **40,** 337–365 (1986).

131. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: Genomic Analysis of Microbial Communities. *Annu. Rev. Genet.* **38,** 525–552 (2004).

132. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345,** 60–63 (1990).

133. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5,** 16–18 (2008).

134. Roesch, L. F. W. *et al.* Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* **1,** 283–290 (2007).

135. Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4,** e368 (2006).

136. Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7,** 57 (2006).

137. Fisher, R. A., Corbet, A. S. & Williams, C. B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *J. Anim. Ecol.* **12,** 42–58 (1943).

138. Mouillot, D. & Leprêtre, A. A comparison of species diversity estimators. *Res. Popul. Ecol.* **41,** 203–215 (1999).

139. Schloss, P. D. *et al.* Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).

140. Chou, H.-H. & Holmes, M. H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17,** 1093–1104 (2001).

141. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **12,** 1889–1898 (2010).

142. Calinski, T. & Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat.* **3,** 1–27 (1974).

143. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5,** e9490 (2010).

144. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).

145. May, A. C. W. Percent sequence identity; the need to be explicit. *Structure* **12,** 737–738 (2004).

146. Schloss, P. D. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71,** 1501–1506 (2005).

147. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5,** 164–166 (1989).

148. Woese, C. R. Bacterial Evolution. *Microbiol. Rev.* **51,** 221–271 (1987).

149. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).

150. Morrison, D. A. & Ellis, J. T. Effects of Nucleotide Sequence Alignment on Phylogeny Estimation: A Case Study of 18S rDNAs of Apicomplexa. *Mol. Biol. Evol.* **14,** 428–441 (1997).

151. Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. & Warnow, T. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324,** 1561–1564 (2009).

152. Krznaric, D. & Levcopoulos, C. Fast Algorithms for Complete Linkage Clustering. *Discrete Comput. Geom.* **19,** 131–145 (1998).

153. Yu, Y., Breitbart, M., McNairnie, P. & Rohwer, F. FastGroupII: a web-based bioinformatics platform for analyses of large 16S rDNA libraries. *BMC Bioinformatics* **7,** 57 (2006).

154. Sun, Y. *et al.* ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.* **37,** e76 (2009).

155. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

156. Milligan, G. W. & Cooper, M. C. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* **50,** 159–179 (1985).

157. Gardner, P. P., Wilm, A. & Washietl, S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.* **33,** 2433–2439 (2005).

158. Larkin, M. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).

159. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

160. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).

161. Engelbrektson, A. *et al.* Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.* **4,** 642–647 (2010).

162. Qu, A. *et al.* Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. *PLoS ONE* **3,** e2945 (2008).

163. Schloss, P. D. A High-Throughput DNA Sequence Aligner for Microbial Ecology Studies. *PLoS ONE* **4,** e8230 (2009).

164. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* **125,** 167–188 (1994).

165. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31,** 3429–3431 (2003).

166. McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29,** 1105–1119 (1990).

167. Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9,** 133–148 (1981).

168. Maulik, U. & Bandyopadhyay, S. Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE T. Pattern Anal.* **24,** 1650–1654 (2002).

169. Tokeshi, M. *Species coexistence: ecological and evolutionary perspectives* (Wiley-Blackwell, New York, 1999).

170. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31,** 343–366 (2000).

171. Hutchinson, G. E. Homage to Santa Rosalia, or why are there so many kinds of animals? *Am. Nat.* **93,** 145–159 (1959).

172. Hutchinson, G. E. The paradox of the plankton. *Am. Nat.* **95,** 137–145 (1961).

173. Chaveé, J., Muller-Landau, H. C. & Levin, S. A. Comparing classical community models: theoretical consequences for patterns of diversity. *Am. Nat.* **159,** 1–23 (2002).

174. Silvertown, J. Plant coexistence and the niche. *Trends Ecol. Evol.* **19,** 605–611 (2004).

175. Wright, S. Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologia* **130,** 1–14 (2002).

176. Caswell, H. Community structure: a neutral model analysis. *Ecol. Monogr.* **46,** 327–354 (1976).

177. Bell, G. The distribution of abundance in neutral communities. *Am. Nat.* **155,** 606–617 (2000).

178. Hubbell, S. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton University Press, Princeton, 2001).

179. Bell, G. *Science* **293,** 2413–2418 (2001).

180. Chave, J. Neutral theory and community ecology. *Ecol. Lett.* **7,** 241–253 (2004).

181. Rosindell, J., Hubbell, S. & Etienne, R. The Unified Neutral Theory of Biodiversity and Biogeography at Age Ten. *Trends Ecol. Evol.* **26,** 340–348 (2011).

182. Muneepeerakul, R. *et al.* Neutral metacommunity models predict fish diversity patterns in Mississippi–Missouri basin. *Nature* **453,** 220–222 (2008).

183. Woodcock, S. *et al.* Neutral assembly of bacterial communities. *FEMS Microbiol. Ecol.* **62,** 171–180 (2007).

184. McGill, B. A test of the unified neutral theory of biodiversity. *Nature* **422,** 881–885 (2003).

185. McGill, B., Maurer, B. & Weiser, M. Empirical evaluation of neutral theory. *Ecology* **87,** 1411–1423 (2006).

186. Hubbell, S. Neutral theory in community ecology and the hypothesis of functional equivalence. *Funct. Ecol.* **19,** 166–172 (2005).

187. Leibold, M. & McPeek, M. Coexistence of the niche and neutral perspectives in community ecology. *Ecology* **87,** 1399–1410 (2006).

188. Purves, D. & L.A., T. Different but equal: the implausible assumption at the heart of neutral theory. *J. Anim. Ecol.* **79,** 1215–1225 (2010).

189. Ricklefs, R. The unified neutral theory of biodiversity: Do the numbers add up? *Ecology* **87,** 1424–1431 (2006).

190. Etienne, R., Alonso, D. & McKane, A. The zero-sum assumption in neutral biodiversity theory. *J. Theor. Biol.* **248,** 522–536 (2007).

191. Allouche, O. & Kadmon, R. A general framework for neutral models of community dynamics. *Ecol. Lett.* **12,** 1287–1297 (2009).

192. Adler, P. B., Rislambers, J. H. & Levine, J. M. A niche for neutrality. *Ecol. Lett.* **10,** 95–104 (2007).

193. Adler, P., Ellner, S. & Levine, J. Coexistence of perennial plants: an embarrassment of niches. *Ecol. Lett.* **13,** 1019–1029 (2010).

194. Volkov, I., J.R., B., S.P., H. & A., M. Inferring species interactions in tropical forests. *Proc. Natl. Acad. Sci. USA* **106,** 13854–13859 (2009).

195. Purves, D., Pacala, S., Burslem, D., Pinard, M. & Hartley, S. in *Biotic interactions in the tropics: their role in the maintenance of species diversity* (eds Burslem, D. F., A., P. M. & E., H. S.) 107–138 (Cambridge University Press, Cambridge, 2005).

196. Chisholm, R. & Pacala, S. Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proc. Natl. Acad. Sci. USA* **107,** 15821–15825 (2010).

197. Gravel, D., Canham, C., Beaudet, M. & Messier, C. Reconciling niche and neutrality: the continuum hypothesis. *Ecol. Lett.* **9,** 399–409 (2006).

198. Tilman, D. Niche tradeoffs, neutrality, and community structure: A stochastic theory of resource competition, invasion, and community assembly. *Proc. Natl. Acad. Sci. USA* **101,** 10854–10861 (2004).

199. Cadotte, M. Concurrent niche and neutral processes in the competition–colonization model of species coexistence. *Proc. R. Soc. B* **274,** 2739–2744 (2007).

200. Zillio, T. & Condit, R. The impact of neutrality, niche differentiation and species input on diversity and abundance distributions. *Oikos* **116,** 931–940 (2007).

201. Loreau, M. & de Mazancourt, C. Species Synchrony and Its Drivers: Neutral and Nonneutral Community Dynamics in Fluctuating Environments. *Amer. Nat.* **172,** 48–66 (2008).

202. Doncaster, C. & Cornell, S. Ecological Equivalence: A Realistic Assumption for Niche Theory as a Testable Alternative to Neutral Theory. *PLoS ONE* **4,** e7460 (2009).

203. Haegeman, B. & Loreau, M. A mathematical synthesis of niche and neutral theories in community ecology. *J. Theor. Biol.* **269,** 150–165 (2011).

204. Dumbrell, A., Nelson, M., Helgason, T., Dytham, C. & Fitter, A. Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J.* **4,** 337–345 (2009).

205. Zhang, Q., Buckling, A. & Godfray, H. Quantifying the relative importance of niches and neutrality for coexistence in a model microbial system. *Funct. Ecol.* **23,** 1139–1147 (2009).

206. Langenheder, S. & Székely, A. J. Species sorting and neutral processes are both important during the initial assembly of bacterial communities. *ISME J.* **5,** 1086–1094 (2011).

207. Ayarza, J. M. & Erijman, L. Balance of neutral and deterministic components in the dynamics of activated sludge floc assembly. *Microb. Ecol.* **61,** 486–495 (2011).

208. Ofiteru, I. D. *et al.* Combined niche and neutral effects in a microbial wastewater treatment community. *Proc. Natl. Acad. Sci. USA* **107,** 15345–15350 (2010).

209. Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S. & Thomas, T. Bacterial community assembly based on functional genes rather than species. *Proc. Natl. Acad. Sci. USA* **108,** 14288–14293 (2011).

210. Horner-Devine, M. C. *et al.* A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* **88,** 1345–1353 (2007).

211. Emerson, B. & Gillespie, R. Phylogenetic analysis of community assembly and structure over space and time. *Trends Ecol. Evol.* **23,** 619–630 (2008).

212. Kelly, C., Bowler, M., Pybus, O. & Harvey, P. Phylogeny, niches, and relative abundance in natural communities. *Ecology* **89,** 962–970 (2008).

213. Cavender-Bares, J., Kozak, K., Fine, P. & Kembel, S. The merging of community ecology and phylogenetic biology. *Ecol. Lett.* **12,** 693–715 (2009).

214. Kembel, S. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26,** 1463–1464 (2010).

215. Cadotte, M. *et al.* Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol. Lett.* **13,** 96–105 (2010).

216. Sipos, M. *et al.* Robust Computational Analysis of rRNA Hypervariable Tag Datasets. *PLoS ONE* **5,** e15220 (2010).

217. Badger, J. H., Ng, P. C. & Venter, J. C. in *Metagenomics of the Human Body* (ed Nelson, K. E.) 1–14 (Springer, New York, 2011).

218. Brulc, J. M. *et al.* Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc. Natl. Acad. Sci. USA* **106,** 1948–1953 (2009).

219. Bäckhed, F., Ley, R., Sonnenburg, J., Peterson, D. & Gordon, J. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–1920 (2005).

220. Dethlefsen, L., McFall-Ngai, M. & Relman, D. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449,** 811–818 (2007).

221. Turnbaugh, P. *et al.* The human microbiome project. *Nature* **449,** 804–810 (2007).

222. Li, M. *et al.* Symbiotic gut microbes modulate human metabolic phenotypes. *Proc. Natl. Acad. Sci. USA* **105,** 2117–2122 (2008).

223. Slack, E. *et al.* Innate and Adaptive Immunity Cooperate Flexibly to Maintain Host-Microbiota Mutualism. *Science* **325,** 617–620 (2009).

224. Antonopoulos, D. *et al.* Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infect. Immun.* **77,** 2367–2375 (2009).

225. Kriegel, H.-P., Kröger, P., Schubert, E. & Zimek, A. in *Scientific and Statistical Database Management* (eds Ludäscher, B. & Mamoulis, N.) 418–435 (Springer, Heidelberg, 2008).

226. Humphray, S. *et al.* A high utility integrated map of the pig genome. *Genome Biol.* **8,** R139 (2007).

227. Muyzer, G., Dewaal, E. & Uitterlinden, A. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* **59,** 695–700 (1993).

228. Adami, C. & Chu, J. Critical and near-critical branching processes. *Phys. Rev. E* **66,** 011907 (2002).

229. Gray, J., Bjorgesaeter, A. & Ugland, K. On plotting species abundance distributions. *J. Anim. Ecol.* **75,** 752–756 (2006).

230. Wang, Q., Garrity, G., Tiedje, J. & Cole, J. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267 (2007).

231. Herrada, E. A. *et al.* Universal scaling in the branching of the tree of life. *PLoS ONE* **3,** e2757 (2008).

232. Herrada, E. A., Eguíluz, V. M., Hernández-García, E. & Duarte, C. M. Scaling properties of protein family phylogenies. *BMC Evol. Biol.* **11,** 155 (2011).

233. Altaba, C. R. Universal artifacts affect the branching of phylogenetic trees, not universal scaling laws. *PLoS ONE* **4,** e4611 (2009).

234. Mayr, E. A natural system of organisms. *Nature* **348,** 491 (1990).

235. Mayr, E. Systems of ordering data. *Biol. Phil.* **10,** 419–434 (1995).

236. Woese, C. R. Default taxonomy: Ernst Mayr's view of the microbial world. *Proc. Natl. Acad. Sci. USA* **95,** 11043–11046 (1998).

237. Blum, M. & François, O. Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance. *Syst. Biol.* **55,** 685–691 (2006).

238. Smrčková, J. *Meta-analysis of methodological artifacts of the phylogenetic imbalance* Master's thesis (University of South Bohemia, 2011).

239. Aldous, D. J. Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Stat. Sci.* **16,** 23–34 (2001).

240. Keller-Schmidt, S., Tuğrul, M., Eguiluz, V. M., Hernandez-Garcia, E. & Klemm, K. An Age Dependent Branching Model for Macroevolution. arXiv:1012.3298 (2010).

241. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26,** 1641–1650 (2009).

242. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37,** D5–D15 (2009).

243. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37,** D26–D31 (2009).

244. Kanz, C. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **33,** D29–33 (2005).

245. Huerta-Cepas, J., Dopazo, J. & Gabaldón, T. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* **11,** 24 (2010).

246. Mir, A., Rosselló, F. & Rotger, L. A new balance index for phylogenetic trees. arXiv:1202.1223 (2012).

247. Brin, S. & Page, L. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN* **30,** 107–117 (1998).

248. Chen, P., Xie, H., Maslov, S. & Redner, S. Finding scientific gems with Google's PageRank algorithm. *J. Informetr.* **1,** 8–15 (2007).

249. Veysey, II, J. J. *Complex fluid dynamics: from laminar to geophysical flows* PhD thesis (University of Illinois at Urbana-Champaign, 2006).

250. in. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (eds Abramowitz, M. & Stegun, I. A.) 875–924 (Dover Publications, New York, 1970).